The optimal design of a criminal justice system[*]

Francisco Silva[†]

October 11, 2015

Abstract

I consider the problem a social planner faces of constructing a criminal justice system that addresses two needs: to protect the innocent and to punish the guilty. I characterize the socially optimal criminal justice system under different assumptions with respect to the social planner's ability to commit. In the optimal system, even before a criminal investigation is initiated, all members of the community are given the opportunity to confess to have committed the crime in exchange for a smaller than socially optimal punishment that is independent of any future evidence that might be discovered. Agents that choose not to confess might be punished once the investigation is completed if the evidence gathered is sufficiently incriminatory. In this paper's framework, exerting leniency towards confessing agents is efficient not because it saves resources or reduces risk, but because there are information externalities to each confession. When an agent credibly confesses to be guilty he indirectly provides the social planner additional information about the other agents: the fact that they are likely to be innocent.

JEL classification: D82, K14

Keywords: leniency, criminal justice system, mechanism design

[*] I would like to thank David Abrams, Mustafa Dogan, Nicholas Janetos, Sangmok Lee, George Mailath, Steven Matthews, Timofyi Mylovanov, Daniel Neuhann, Qiusha Peng, Andrew Postlewaite, Rakesh Vohra and Yanhao Wei as well as the UPenn's Micro Lunch seminar participants for their useful comments. This is a preliminary version of my job market paper. I have recently become aware of a paper by Siegel and Stulovici that is closely related to this paper. I have not had a chance to revise this version to discuss that paper but will do so in future versions of this paper.

[†] Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104 (email: fsilva@sas.upenn.edu).

1 Introduction

In this paper, I investigate the desirable properties of a criminal justice system. I approach this problem by analyzing a simple scenario that I believe illustrates the main challenge a criminal justice system faces. Consider a community of N agents and a principal. Imagine that the suspicion that a crime has been committed arises and it is the principal's responsibility to select punishments to be inflicted upon the agents. Think of the principal as a sort of social planner or benevolent decision maker who wants the best for the community. In a perfect world, she would would like to punish only the agents that were guilty of committing the crime, while leaving the innocent ones unharmed. 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 that are guilty makes these reluctant to announce their guilt. I study the principal's problem of creating a mechanism that, to the extent that is possible, is able to punish those that are guilty in an appropriate manner and still protect the rights of the innocents.

There are two particular ways to solve this problem - two systems - that are important for my analysis. The first one is what I call a "trial" system. Throughout the history of civilization, the traditional way societies have tried to solve this problem has been through a trial system. If the suspicion that a crime has been committed arises, the principal initiates a police investigation aimed at obtaining evidence. Based on such evidence, the principal forms beliefs about the guilt of each of the agents and chooses punishments accordingly. In particular, only agents whose evidence is strongly indicative of guilt are punished - agents are punished if they are found to be guilty beyond "reasonable doubt". The merits of this system come from the fact that the evidence is more likely to point to guilt if the agent is indeed guilty than if he is not. In this paper, I argue that, in general, trial systems are not optimal.

The second important system is what I call a "confession inducing system" (CIS). A CIS is defined to have two stages. In the first stage, even before an investigation is initiated, all N agents are given the opportunity to confess the crime if they so choose, in exchange for a constant punishment, independent of any evidence that might be gathered in the future and that only depends on the nature of the crime. In the second stage, if necessary, the principal conducts a police investigation in order to collect evidence about the crime, and, based on the information gathered, chooses the punishments, if any, to apply to all agents who chose not to confess in the first stage. It essentially represents a trial system that is preceded by a confession 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 with self-reporting is that firms that infringe some environmental regulation are able to contact the law enforcement authorities 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 similar system is plea bargaining in criminal law where defendants are given the chance to confess to have committed the crime in exchange for a reduced sentence.

There are two main results in this paper that, I believe, make a contribution to the understanding of the desirable properties of a criminal justice system. First, in general, it is possible to construct a CIS that is strictly preferred to any trial system and second, in general, it is possible to construct a CIS that is preferred to any other system, no matter how complicated it might be, and under different assumptions with respect to the principal's ability to commit.

The first result, in it of itself, should not come as a surprise for most readers. In the United States, the percentage of criminal cases resolved through Plea Bargaining is as high as 97% (Dervan and Edkins (2013)), which clearly indicates that Plea Bargaining has become the norm rather than the exception. The main justification for this widespread use of Plea Bargaining is that it is just not feasible to grant every defendant a criminal trial. There are simply not enough resources to pay all the lawyers, judges and jurors that a trial requires. And so, the only feasible solution is Plea Bargaining.[1] In a paper of 1994, Kaplow and Shavell argue that it is possible to build a CIS that is as good as any trial system, with the added benefit of being cheaper. I believe one interesting aspect of my paper is that I argue that there are CISs that are preferred to trial systems, not only because they are cheaper, but because they are better at appropriately punishing guilty agents, while preserving the rights of the innocents.

In this paper, I identify two main virtues of CISs, which are more easily understood through a very simple example. Imagine that, in a small town, there has been a big fire that has damaged the local forest. From the moment the principal witnessed the fire she became suspicious that it might not have been an accident. In the meantime, she has done some investigative work already and was lucky enough to narrow down her list of suspects to a single agent - agent 1. However, she remains unsure of whether the agent is indeed guilty, or if the fire was simply an accident. As a result, she has requested that a modern device be sent to her from a different country that would allow for the analysis of the residues collected from the forest, which might shed light on what has happened. The device is supposed to arrive in a few days.

Let the continuous random variable $\theta_1 \in [0,1]$ represent the evidence collected from analyzing the residues and assume that larger values of θ_1 are relatively more likely to have been generated if agent 1 is guilty rather than if he is innocent. Formally, assume $\frac{\pi(\theta_1|t_1=g)}{\pi(\theta_1|t_1=i)}$ is strictly increasing, where $\pi(\theta_1|t_1)$ represents the probability density function of θ_1 conditional on agent 1 being either guilty ($t_1 = g$) or innocent ($t_1 = i$). This means that the larger θ_1 is, the more likely it is that the fire was not an accident, and that agent 1 is guilty. For example, if the principal is able to identify agent 1's footprint from the collected residues, then θ_1 should be large.

In a trial system, the principal is supposed to wait for the new device to arrive, collect and analyze the residues (i.e. observe θ_1), form beliefs about the guilt of agent 1 and then choose whether to punish him. In particular, it seems natural to expect that, in such a system, the agent receives some normalized punishment of 1 if the principal is sufficiently convinced he is guilty, and is acquitted otherwise. Therefore, there is going to be a threshold $\overline{\theta}_1$ such that the agent is convicted if and only if $\theta_1 > \overline{\theta}_1$ - see Figure 1. This threshold $\overline{\theta}_1$ is endogenous and represents the standard of proof the principal uses to make his decision. It depends very much on how large the principal's concern about wrongly punishing innocent agents is.

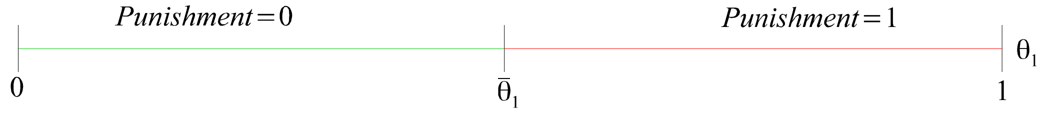

Figure 1: The trial system

For concreteness, assume that $\pi\left(\theta_1 > \overline{\theta}_1 | t_1 = g\right) = \frac{3}{4}$ and $\pi\left(\theta_1 > \overline{\theta}_1 | t_1 = i\right) = \frac{1}{2}$, which implies that the expected punishment agent 1 receives, conditional on him being guilty (denoted by B_1^g) is equal to $\frac{3}{4}$, and conditional on him being innocent (denoted by B_1^i) is equal to $\frac{1}{2}$.

Now, assume that the agent is risk neutral and that the principal has commitment power, and consider the following alternative. Imagine that, before the new device arrives, the principal approaches agent 1 and gives him the opportunity to confess to be guilty in exchange for a constant punishment of $\frac{3}{4}$. If the agent refuses, then everything is as before - the principal waits for the device to arrive and punishes the agent in 1 if and only $\theta_1 > \overline{\theta}_1$. The punishment of

[1] The United States Supreme Court has explicitly encouraged this practice, for example, in Santobello v. New York (1971), precisely on these grounds.

$\frac{3}{4}$ is chosen exactly to make the agent indifferent when guilty, giving him just enough incentives to confess, while if he is innocent he prefers not to. Therefore, in this alternative CIS, agent 1's expected punishment is the same as in the trial system regardless of whether he is innocent or guilty. This equivalence is what led Kaplow and Shavell (1994) to argue for the superiority of CISs with respect to trial systems on the grounds that the latter uses less resources - if agent 1 confesses the crime there is no need to collect evidence anymore. In this paper, however, I assume there are no costs of any nature and so these two systems are considered equivalent.

Now, I show that it is possible to create a new CIS that reduces the expected punishment of agent 1 when he is innocent (reduces B_1^i) while keeping it constant and equal to $\frac{3}{4}$ when he is guilty ($B_1^g = \frac{3}{4}$). I propose to do this by increasing the standard of proof from $\overline{\theta}_1$ to $\widehat{\theta}_1$, where $\widehat{\theta}_1$ is such that $\pi\left(\theta_1 > \widehat{\theta}_1 | t_1 = g\right) = \frac{1}{2}$ and $\pi\left(\theta_1 > \widehat{\theta}_1 | t_1 = i\right) = \frac{1}{4}$, so that if the agent chooses not to confess, he is less likely to be punished. The problem with this new CIS is that, when agent 1 is guilty, he no longer prefers to confess. So, one must increase the second stage punishment just enough, in order to provide him with enough incentives to confess. It follows that, if $\theta_1 > \widehat{\theta}_1$, the agent should receive a punishment of $\frac{3}{2}$ (because $\frac{3}{2} * \frac{1}{2} = \frac{3}{4}$) if he has not confessed in the first stage - see Figure 2.

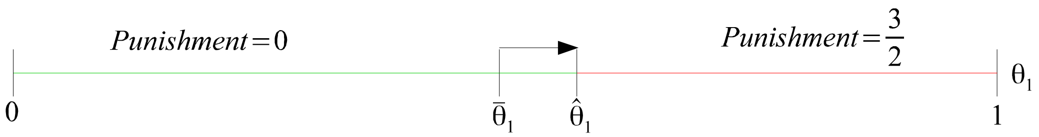

Figure 2: Second stage punishments of the new CIS

In this new CIS, $B_n^g = \frac{3}{4}$ because the agent is confessing the crime when guilty, but $B_n^i = \frac{3}{8} < \frac{1}{2}$ - the agent is made better off only when innocent. This happens because of the monotone likelihood ratio property of θ_1. When one increases the threshold from $\overline{\theta}_1$ to $\widehat{\theta}_1$, the relative impact of this change is higher if the agent is innocent than if he is guilty. In particular, the probability of conviction if the agent is innocent is reduced by 50% (from $\frac{1}{2}$ to $\frac{1}{4}$), while if the agent is guilty it is only reduced by 33% (from $\frac{3}{4}$ to $\frac{1}{2}$). Therefore, when the second stage punishment is increased to make the agent indifferent when guilty, it ends up being small enough for the agent to be made better off when innocent.

Notice that this method is only possible if the principal is allowed to "overpunish", i.e. to punish an agent in more than the maximum punishment admissible in the trial system. However, it seems questionable to me whether it is desirable and even possible to construct a system that enforces arbitrarily large punishments. Take as an example the crime of arson. If the fire in question did not injure anyone and only caused material damage, it does not seem reasonable to me to expect that a system that inflicts a punishment of, say, 50 years of imprisonment or worse to the agent is going to be accepted by society. This is even more true for crimes of lesser importance, like minor theft. Suppose one does not allow the principal to overpunish and imposes an upper bound of 1 to all punishments. Is it still the case that there are CISs that are strictly preferred to any trial system? In general, the answer is *yes*, provided $N \geq 2$.

Consider the same arson example, but at an earlier stage. In particular, imagine the principal has just witnessed the fire. At this moment, the principal cannot rule out anyone from the community as being guilty as she has yet to collect any evidence. She simply believes that there is some probability a crime has been committed and that each agent in the community might be guilty.

If the principal implements a trial system, she is supposed to collect all the available evidence, and then, based on it, choose how to punish all agents. Consider the following alternative CIS. Imagine that, before initiating the investigation, the principal gives agent 1 the opportunity to confess in exchange for a constant punishment that leaves him indifferent only if guilty. After agent 1 has chosen to either confess or not, then the principal initiates an investigation aimed at producing evidence which is then used to select the punishments of all other agents (as well as agent 1 if he chose not to confess). As described above, in this new mechanism, agent 1 only confesses when guilty, and his expected punishment is kept intact, regardless of whether he is guilty or not. But now consider what happens to the remaining agents. When judging each of the other agents the principal will still have collected the same evidence as with the trial system, but now, will also be informed of whether agent 1 is guilty or innocent - he is guilty if he chose to confess and innocent otherwise. This means that the decision the principal makes with respect to any of the other agents is more accurate, as he has more relevant information. For example, if agent 1 confesses, the principal should be very certain that the other agents are innocent, and so is less likely to make a mistake. In other words, there are information externalities in an agent's confession. By reporting to be guilty an agent is not only making a statement regarding his own guilt, but he is also saying that the other agents are likely to be innocent. Even though this is not the optimal CIS - in the optimal CIS every agent is given the opportunity to confess - it illustrates the shortcomings of the trial system and highlights the potential benefits of allowing agents to confess to have committed the crime before an investigation has been initiated.

Implicit in this argument is that the information each agent holds (whether they are innocent or guilty) is important in evaluating others' guilt - the agents' innocence is correlated. This assumption is usually well accepted for a certain set of crimes that are likely to be committed by an organized group - for example, in anti-collusion legislation, because there is the sense that each cartel member is likely to have information about the other cartel members, it is often possible for them to confess their guilt in exchange for a smaller punishment. What this simple example about a fire illustrates is that the same argument should be used for the majority of the "normal" crimes, because, in each of these, the knowledge that a given agent is guilty is likely to reveal the innocence of others.

Notice also that, even though this argument requires $N \geq 2$, this does not mean that it only applies to criminal cases where there are multiple defendants. Consider again the arson example. The principal first becomes suspicious that a crime may have been committed when she first witnesses the fire. At this moment, she is likely to know very little about who the criminals might be and so, everyone is a suspect, in the sense that no agent has yet been ruled out. So, by applying my analysis to that moment of time, it is clear that $N \geq 2$, as N should include everyone that the principal believes might, with some probability, have committed the crime. Hence, the interpretation of what constitutes an "agent" depends very much on the context. In particular, it only makes sense to think of an agent as a defendant if one is analyzing the later stages of the criminal process. However, my analysis suggests that the opportunity to confess should be given as early as possible, and not later when some evidence has already been gathered, because confessions are easier to induce (a guilty agent is more afraid that future evidence might incriminate him) and provide more information to the principal. It then follows that an optimal system would give agents the opportunity to confess as soon the crime has been committed, through, for example, self-reporting legislation, where each person is always able to confess to have committed a specific crime and receive a pre-determined punishment as a consequence.

The second important result of this paper is that, in general, there is a CIS that is optimal for the principal. To the best of my knowledge, this is one of the first papers in the law enforcement literature that applies mechanism design techniques to the study of the desirable properties of a criminal justice system. Using these type of techniques it is possible to compare the CIS not only to the trial system but to any other system, no matter how complicated it might be. From a theoretical point of view, I believe this result is important because it renders the search for a better system unnecessary, at least within the context of my model. From a policy point of view, this result is also important because of the simplicity of CISs. The only requirement for its implementation is the guarantee by the principal that agents are given the opportunity to confess in exchange for a constant punishment.

In the first part of the paper, I show these two results while assuming the principal has commitment power. As it is clear from the fire example, this assumption is important as it allows the principal to i) impose smaller than optimal punishments on knowingly guilty agents (the ones that confess) and ii) punish knowingly innocent agents (the ones that do not confess). In the second part of the paper, I relax this assumption in two ways.

First, I consider the class of renegotiation proof mechanisms, where only i) is permitted. The idea is that if the principal is supposed to punish an agent he knows is innocent, both her and the agent would have an incentive to renegotiate such punishment, as they would both prefer a smaller one. In this setup, I show the two results still hold - it is possible to construct a CIS that is strictly preferred to the trial system and (maybe weakly) preferred to any other system. Second, 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 it is not possible to improve upon the trial system.

The structure of the paper is as follows. In section 2, I analyze the related literature. In section 3, I present the model. In section 4, as a benchmark, I formalize the trial system. In section 5, I analyze the second best problem - I look for a Bayes-Nash incentive compatible allocation that maximizes the principal's utility when the agents' innocence is private information and the principal has commitment power. In section 6, I restrict the set of possible allocations to the ones that can be implemented through a) a renegotiation proof mechanism and b) a sequentially optimal mechanism. In section 7, I consider four extensions to the model. In the first one, I allow for risk averse agents and show that CISs are still optimal even when innocent agents are more risk averse than guilty agents. In the second extension, I allow for a richer information structure that takes into account the fact that guilty agents might be a part of a conspiracy. In the third extension, I allow for some additional privately observed heterogeneity among the agents. And, finally, in the fourth extension, I consider a change in the timing of the model and assume the principal is only able to propose a mechanism after gaining knowledge about the evidence. In section 8, I conclude.

2 Related Literature

There is a considerable amount of literature in economics that argues for the use of variants of CISs 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 the social welfare because it saves monitoring costs. By setting the punishment upon 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 merits of plea bargaining. In particular, they point out that plea bargaining reduces the risk that exists at trial of acquitting guilty agents. The authors assume that, if an agent goes to trial, he is either punished to the extent of the crime or acquitted, depending on the evidence. The argument is that it is better to let the guilty agents confess and punish them with the corresponding certainty equivalent punishment (a constant punishment that makes guilty agents indifferent between confessing the crime and going to the trial). If the principal's preferences are such that he wishes to minimize the risk associated with a wrongful acquittal, he will prefer this outcome even if the certainty equivalent is equal to the expected punishment of the guilty agent at trial (in case the agent is risk neutral). In fact, even if the principal does not have risk concerns but the agent is risk averse, then it follows that the certainty equivalent punishment is larger than the expected punishment at trial for the guilty agent, which would also mean that the principal would be better off.

This paper builds on these in that it highlights an additional advantage of CISs: by inducing guilty agents to confess to have committed the crime, the principal obtains relevant information when assessing the guilt of other agents.[2]

A feature common to both 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 my model, the main problem about assuming the principal has commitment power is the fact that it allows her to punish agents that she knows are innocent. This is problematic because, in such an event, both the principal and the agent would have an incentive to renegotiate that punishment and agree on eliminating it. In section 6, I study the problem of finding the optimal renegotiation proof mechanism and argue that CISs are still optimal despite not achieving full separation between innocent and guilty agents, which means that, in equilibrium, only a fraction of the guilty agents confesses to have committed the crime.

A key aspect of my argument has to do with the fact that the principal deals with different agents whose types (their innocence) may not be independent. There are a few articles on law enforcement that have also considered multiple defendants, but the emphasis is not on distinguishing the innocent agents from the guilty ones, but rather to find the optimal strategy in order to achieve maximum punishment for the defendants, as they are all assumed to be guilty - for example Kim (2009), Bar-Gill and Ben Shahar (2009) and Kobayashi (1992). There is also a literature on industrial organization that considers the design of leniency programs in Antitrust law that also considers multiple agents - see Spagnolo (2006) for a literature review.

In mechanism design, there is a literature that analyzes how the optimal mechanism depends on the correlation between the agents' types. Cremer and McLean (1988) show that, if there is correlation, it is generally possible to implement the efficient allocation. The idea is that it is possible to construct a lottery of payments for each type of agent that is appealing for that type but unappealing for others. This difference in how appealing the same lottery is exists because each type has different beliefs about the other agents' types. In section 5, I show that, in my model, it is possible to implement the first best solution - innocent agents are acquitted and guilty agents are not. The principal is able to accomplish this not by using the correlation between the agents' types but because of the way the evidence is generated. In particular, if the principal selects to punish only when the evidence is very likely to have been generated by a guilty agent and, in that event, imposes a very large punishment, she is able to construct a lottery of punishments at trial that simultaneously imposes an expected punishment close to 0 if the agent is innocent but close to 1 if he is guilty. However, this method is clearly not possible when the principal's commitment power is restricted, as she would always prefer to

[2] Grossman and Katz (1983) mention a related effect associated with plea bargaining that they call "screening effect" - given that only guilty agents plead guilty, the prosecutor is able to distinguish them from the innocent agents. However, such distinction ends up being irrelevant in their model as this effect has no welfare impact when there is only one agent (as I show in section 5). Even though the guilty agents are identified, they are still punished as harshly as they would have been if there was no interaction between them and the principal. The only welfare effect that exists in the environment of Grossman and Katz (1983) is due to the relation with risk that both the principal and the agents have.

reduce such punishments. And even when the principal does have commitment power, it is neither feasible nor desirable - presumably the highest punishment one can inflict is the death penalty and even that is only accepted in very rare occasions.

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 generally as "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 that considers similar environments by assuming 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 distinguishing 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 bigger 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 that are wealth-constrained. Another approach is to assume 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 is costly but 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.[3]

3 Model

There are N agents and a principal. Each agent n randomly draws a type $t_n \in \{i, g\} \equiv T_n$ that is his private information - each agent n is either innocent (i) or guilty (g) of committing the crime. Let $T = \{T_n\}_{n=1}^{N}$ be the set of all possible vectors of agents' types and $T_{-n} = \{T_j\}_{j \neq n}$ 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 assumed to be common knowledge.

This description implicitly assumes each agent knows only whether he is innocent or guilty, and has no other relevant information about other agents' innocence. Thus, it rules out crimes that 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.

There are two reasons for me to have made this assumption. First, the majority of crimes are not conspiracy crimes. In most crimes, by their nature, it is very unlikely that more than a single individual has committed the crime. And there are even crimes that, although being committed by several agents, are not conspiracy crimes in the sense that it is unlikely the agents know each other - think, for example, of tax evasion. Because the simple model is already broad enough to analyze most crimes, I did not think it would be beneficial to generalize the model at the expense of its simplicity. The second reason is that the main intuitions still hold true if I was to consider a more complicated information structure, as becomes clear in section 5. Because of these two reasons, I have decided to analyze more closely the case of conspiracy crimes in one of the extensions of this paper, in section 7.2. In it, I generalize the model and show that an "extended" CIS, where each agent is given the opportunity to incriminate other

[3] 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.

agents when confessing, is optimal and further discuss what the agents' punishment should depend upon.

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 $\Theta = \prod_{n=1}^{N} \Theta_n = [0,1]^N$ denote the set of all possible evidence vectors, while $\Theta_{-n} = [0,1]^{N-1}$ denotes the set of all possible evidence vectors that exclude only agent n's evidence level. The evidence vector θ is made of exogenous 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 $\pi(\theta_n|t_n)$, while the joint conditional pdf of θ given t is denoted by $\pi(\theta|t) = \prod_{n=1}^{N} \pi(\theta_n|t_n)$. (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 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 fingerprint of agent n is not found, this information is still contained in θ_n. Also, the assumption of conditional independence of $\theta_n|t_n$ is mostly made for simplicity as no result depends on it. In particular, notice that it does not imply that θ_n is independent of θ_{-n}.

Let $l(\theta_n) = \frac{\pi(\theta_n|t_n=g)}{\pi(\theta_n|t_n=i)}$ be the evidence likelihood ratio. I assume 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 $\lim_{\theta_n \to 0} l(\theta_n) = 0$ and $\lim_{\theta_n \to 1} l(\theta_n) = \infty$, 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.

I assume agent n's utility is given by $u^a(x_n) = -x_n$ where $x_n \in \mathbb{R}_+$ represents the punishment he 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) (which I discuss in the related literature section), where the advantage of CISs relative to trial systems comes from the fact that agents are risk averse. In one of the extensions, in section 7.1, I analyze the case where agents are allowed to be risk averse and show that CISs are still optimal, even when innocent agents are more risk averse than guilty ones.

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. Her utility depends not only on the punishment she inflicts but also on whether the agent that receives it is innocent or guilty. I assume the principal's utility function is given by $u^p(t,x) = \sum_{n=1}^{N} u_n^p(t_n, x_n)$ for all $t \in T$ and $x = (x_1, .., x_N) \in \mathbb{R}_+^N$, where $u_n^p(t_n, x_n) = \begin{cases} -\alpha x_n \text{ if } t_n = i \\ -|1 - x_n| \text{ if } t_n = g \end{cases}$ with $\alpha > 0$. 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, that I normalize to 1. In either case, deviations from the preferred punishments induce a linear cost on the principal.[4] This punishment of 1 that "fits the crime" is exogenous to the model and is likely to be influenced by the nature of the

[4] Grossman and Katz (1983) also assume that there is a punishment that fits the crime. The only difference is that they assume a strictly concave cost upon deviations rather than a linear one. An alternative assumption would be to have the principal simply maximize the punishment imposed on guilty agents rather than having a bliss punishment, in which case my main results would still hold.

crime - the punishment that fits the crime of murder is larger than the punishment that fits the crime of minor theft. As it will be become clear in section 4, this will be the punishment imposed at trial when the agent is found guilty. 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 agent and is small if she is more concerned with wrongly acquitting guilty agents.

Notice that, at first blush, it might appear as though the assumed principal's preferences are too restrictive in that they ignore one of the most important goals of any criminal justice system - to deter crime. In particular, if the goal of the principal is to deter crime, she should want to maximize $\left\{B_n^g - B_n^i\right\}$ - the difference between the expected punishment when the agent is guilty and when he is innocent. In section 5, I address this observation in detail and argue that these deterring preferences can be thought of as a special case of the preferences I have assumed by considering a particular α that is chosen in an appropriate way.[5]

Finally, notice that, under complete information and for any α, the first best allocation $x^{FB} = \left(x_1^{FB}, ..., x_N^{FB}\right)$ is given by

$$x_n^{FB} = \begin{cases} 1 \text{ if } t_n = g \\ 0 \text{ if } t_n = i \end{cases} \quad \text{for all } n$$

4 Trial System

I define the trial system as one where there is no communication between the principal and the agents. The principal simply makes punishment decisions after having collected all the evidence, and imposes those punishments on the agents, that do not have any active role. Let $X^{Tr} = \left\{x : \Theta \rightarrow \mathbb{R}_+^N\right\}$ be the set of possible allocations that are implementable through a trial system. The principal will choose an allocation from X^{Tr} in order to maximize his ex-ante expected utility that is given by

$$V(x) = \int\limits_{\theta \in \Theta} \sum_{t \in T} \pi(t, \theta)\, u^p(t, x)\, d\theta$$

where $\pi(t, \theta) = \pi(\theta | t)\, \pi(t)$.

Notice that we can write $V(x) = \sum\limits_{n=1}^{N} V_n(x_n)$ where

$$V_n(x_n) = \int\limits_{\theta \in \Theta} \sum_{t \in T} \pi(t, \theta)\, u_n^p(t_n, x_n)\, d\theta$$

Therefore, it follows that the choice of the optimal $x \in X^{Tr}$ consists of N independent choices of x_n that each maximize $V_n(x_n)$. Realizing that a punishment higher than 1 is not optimal and further simplifications allows for writing $V_n(x_n)$ as

$$\int\limits_{\theta \in \Theta} \left(\pi(t_n = g | \theta) - \alpha \pi(t_n = i | \theta)\right) \pi(\theta)\, x_n(\theta)\, d\theta - k \tag{1}$$

where k is some constant, independent of x_n, $\pi(\theta) = \sum\limits_{t \in T} \pi(t, \theta)$ for all $\theta \in \Theta$ represents the marginal pdf of θ and $\pi(t_n | \theta)$ is the conditional probability of agent n being of type t_n given the realized evidence vector θ.

[5] See Figure 4 and the subsequent discussion.

Condition (1) displays the simple basis for the principal's decision in a trial system. If $\pi(t_n = g|\theta) \geq \alpha\pi(t_n = i|\theta)$, the principal is convinced enough that agent n is likely to be guilty, given the evidence presented, and will prefer to inflict a punishment of 1 upon him. If not, the principal believes agent n is likely to be innocent, and will acquit him. In this context, the parameter α is a measure of the standard of proof - if α is large, the evidence must be largely indicative of guilt for the agent to be convicted.

Denote the optimal trial solution by x^{Tr}. Given the monotone likelihood ratio property assumed on the evidence, it is possible to describe x^{Tr} as

$$x_n^{Tr}(\theta) = \begin{cases} 1 \text{ if } \theta_n > \theta_n^{Tr}(\theta_{-n}) \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } n$$

where $\theta_n^{Tr}(\theta_{-n})$ is completely characterized in Proposition 1. The principal follows a threshold rule where he convicts the agent if and only if his evidence level θ_n is above such threshold.

Proposition 1 $\theta_n^{Tr}(\theta_{-n}) = l^{-1}\left(\alpha \frac{\sum\limits_{t_{-n}\in T_{-n}} \pi(i,t_{-n}) \prod\limits_{\tilde{n}\neq n} \pi(\theta_{\tilde{n}}|t_{\tilde{n}})}{\sum\limits_{t_{-n}\in T_{-n}} \pi(g,t_{-n}) \prod\limits_{\tilde{n}\neq n} \pi(\theta_{\tilde{n}}|t_{\tilde{n}})}\right)$

Proof. See appendix. ∎

The threshold $\theta_n^{Tr}(\theta_{-n})$ depends on θ_{-n} and so the decision about the conviction/acquittal of agent n is not independent of the evidence of other agents. This is because agents' types may be correlated (which means that information about other agents' types is useful for the principal's decision) and each agent's evidence level is informative of that agent's guilt.

5 Second Best Problem

In this section, I analyze the problem the principal faces of constructing an optimal system, under the assumption that she has commitment power. I assume that, before any evidence is generated, but after agents have gained knowledge of their own type, the principal proposes a mechanism. So, in terms of the example, I analyze the principal's problem when she first witnesses the fire, and has yet to gather any evidence.[6] From 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.

In this context, an allocation is a mapping from the agents' types and their evidence level to the punishments that each of them will be given. Let $X^{SB} = \left\{x : T \times \Theta \to \mathbb{R}_+^N\right\}$ be the set of all such allocations. An allocation $x \in X^{SB}$ is (Bayes-Nash) incentive compatible if and only if, for all $t_n \in T_n$, for all $t_{-n} \in T_{-n}$ and for all n,

$$-\int\limits_{\theta\in\Theta} \sum_{t\in T} \pi(t,\theta|t_n)\, x_n(t_n, t_{-n}, \theta)\, d\theta \geq -\int\limits_{\theta\in\Theta} \sum_{t\in T} \pi(t,\theta|t_n)\, x_n(t_n', t_{-n}, \theta)\, d\theta \text{ for all } t_n' \in T_n \tag{IC}$$

where $\pi(t,\theta|t_n)$ represents the conditional joint pdf of (t,θ), given t_n.

The condition states that, prior to the discovery of the evidence and given allocation x, the expected utility of type t_n of agent n is higher if he reports truthfully than if he misreports, when all other agents are also reporting truthfully.

[6] In one of the extensions, in section $7,4$, I consider a different time frame where the principal is privately informed of the evidence, and only then proposes a mechanism.

I impose an additional condition on the incentive compatible allocations: an upper bound of $\phi \geq 1$ on each punishment:

$$x_n (t, \theta) \leq \phi \text{ for all } t, \theta \text{ and for all } n \tag{UB}$$

The reason for this upper bound is to complement the principal's preferences stated above. What the condition means is that it is so undesirable for a society to punish agents too harshly that it just will not allow it. Imagine the crime that one is referring to is theft and that society finds that a one year of imprisonment is the appropriate punishment for guilty agents. It is not reasonable to expect that society will accept that any agent accused of theft ends up convicted by, say, ten years. In fact, an argument can be made that the highest punishment a society is willing to accept in such cases is exactly one year. With this last observation in mind, I give special attention to the case of $\phi = 1$ below.

The problem I wish to solve is to select an allocation from X^{SB} that maximizes V subject to (IC) and (UB). As in the previous section, because it is possible to write $V(x) = \sum\limits_{n=1} V_n (x_n)$, the problem of finding the optimal allocation can be made into N independent problems where, for each n, x_n is chosen to maximize $V_n (x_n)$ subject to agent n's incentive and upper bound constraints.

There are two earlier remarks that are important to characterize the optimal allocation. First, the innocent's incentive constraint does not bind and, therefore, can be disregarded. To see this, consider the problem where the innocent's incentive constraint is disregarded. The solution of that problem must still satisfy the disregarded incentive constraint for if it did not, the principal could set the punishments imposed upon innocent reports to equal the ones of the guilty reports. This new allocation would be incentive compatible (as it would not depend on the agent's own report) and would improve the expected utility of the principal (because it would decrease the expected punishment of the innocent agents).

Second, punishments imposed on guilty agents never exceed 1. Increasing the punishments on guilty agents to more than 1 decreases the principal's expected utility and does not give more incentives for guilty agents to report truthfully, quite the opposite.

These two remarks allow V_n to be written as

$$\pi (t_n = g) B_n^g - \alpha \pi (t_n = i) B_n^i - k \tag{2}$$

where $\pi (t_n) = \sum\limits_{t_{-n} \in T_{-n}} \pi (t_n, t_{-n})$ is the probability that agent n is of type t_n and $B_n^{t_n}$ represents the expected punishment of agent n, when is of type t_n.

The remaining incentive constraint - when agent n is guilty - can be written as

$$B_n^g \leq \int\limits_{\theta \in \Theta} \sum_{t_{-n} \in T_{-n}} \frac{\pi (g, t_{-n}, \theta)}{\pi (t_n = g)} x_n (i, t_{-n}, \theta) \, d\theta \tag{3}$$

From (2) and (3), it follows that it is optimal to set $x_n (g, t_{-n}, \theta) = B_n^g$ for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$ - if the agent is guilty, he is to receive a constant punishment. This is because both (2) and (3) only depend on B_n^g and not on how the guilty punishments are distributed.

There is one last remark that simplifies the problem. In any solution, the guilty agent is indifferent between reporting his guilt and lying and reporting to be innocent. The reason is that if he is not indifferent and strictly prefers to report truthfully, the principal could reduce the punishments imposed upon innocent reports and still have an incentive compatible allocation. This change would be beneficial for the principal as it would reduce the expected punishment of the innocent agents that are reporting truthfully. Therefore, in an optimal solution, (3) must hold with equality. By plugging (3) into (2), it is possible to write the new objective function of the principal solely as a function of the punishments to be imposed on the innocent type. In particular, the principal's *simplified* nth agent problem is to choose $x_n(i, t_{-n}, \theta) \in [0, \phi]$ for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$, in order to maximize

$$\int_{\theta \in \Theta} \sum_{t_{-n} \in T_{-n}} \left(\pi(g, t_{-n}, \theta) - \pi(i, t_{-n}, \theta)\right) x_n(i, t_{-n}, \theta) \, d\theta - k \tag{4}$$

subject to

$$\int_{\theta \in \Theta} \sum_{t_{-n} \in T_{-n}} \frac{\pi(g, t_{-n}, \theta)}{\pi(t_n = g)} x_n(i, t_{-n}, \theta) \, d\theta \leq 1 \tag{5}$$

Condition (5) simply states that B_n^g, which is equal to the left hand side of (5) by (3), does not exceed 1 given that it is not optimal to overpunish guilty agents.

The case of $\phi = 1$

I believe the case of $\phi = 1$ deserves special attention. If $\phi > 1$, this means that it is possible for the principal to impose punishments that are above what she would deem appropriate if she knew the agent was guilty. As I discuss in more detail below, the principal will be able to use this ability to overpunish in order to improve the quality of the allocation. However, it is highly debatable whether the principal is (or should be) able to impose such high punishments. This practice is reminiscent of alleged prosecutor strategies of inflating the severity of the accusations to persuade defendants to accept plea deals in criminal cases. Such a practice has been widely condemned (see White (1979) or Scott and Stuntz (1992)) precisely on the basis that punishments above what are deemed appropriate are not acceptable.

If $\phi = 1$, constraint (5) does not bind. This is because, if all innocent punishments are bounded by 1, its weighted average must also be bounded by 1.

Therefore, it follows directly from (4) that the optimal punishment to be inflicted upon an innocent agent is 1 if

$$\pi(t_n = g | t_{-n}, \theta) > \alpha \pi(t_n = i | t_{-n}, \theta) \tag{6}$$

and 0 otherwise, where, for simplicity, I assume ties are broken in favor of an acquittal.

As for the punishments to be imposed on guilty agents, the only condition necessary is that the expected punishment of a guilty agent leaves him indifferent to misreporting. If $\phi = 1$, there are several allocations that accomplish this. The particular allocation this paper is interested in is one where, if an agent reports to be guilty, he receives a constant punishment. This allocation is important because it can be implemented by a CIS as follows. In the first stage, all agents are simultaneously given the opportunity to confess. If agent n confesses, he is to receive a constant punishment of $B_n^g \in [0, 1]$. If he refuses, he proceeds to the second stage where he is to be punished according to condition (6). The optimal allocation is implemented by having guilty agents confess and innocent agents not to.

Proposition 2 *If $\phi = 1$, there is a CIS that implements a second best optimal allocation.*

CISs are appealing, within the set of optimal systems, for a number of reasons. First, they are simple. The only requirement is that each agent has the opportunity to confess the crime, which means that the majority of agents, who are likely to be innocent, have a passive role in the system. Second, they are cheaper. In a CIS, if an agent confesses, his punishment is independent of the evidence that might be collected, unlike in any other optimal system. This means that the costs of collecting and analyzing the evidence are reduced. And finally, variants of CISs already exist under a variety of forms, like plea bargaining in criminal law and self-reporting regulation in environmental law.

Recall that, in a trial system, an agent has no other choice but to go to trial and be punished if $\pi\left(t_n = g|\theta\right) > \alpha\pi\left(t_n = i|\theta\right)$, i.e. if, given the evidence, the principal believes he is likely to be guilty. In a CIS an agent may choose whether to go to (the second stage) trial or not. If he chooses to go to trial, he is punished if $\pi\left(t_n = g|t_{-n}, \theta\right) > \alpha\pi\left(t_n = i|t_{-n}, \theta\right)$. This means that the second stage trial that is a part of the CIS is more accurate than the trial system. While in the trial system, the principal only uses the evidence gathered to evaluate the guilt of the agent, in a CIS, in addition to the evidence, the principal is informed of whether other agents are guilty. This information is, in general, relevant, as knowing that one agent is guilty makes it more likely that the remaining agents are innocent. If all agents actually chose to go to the second stage trial in the CIS, this observation would be enough to find it strictly preferred to the trial system. But that is not the case as, in equilibrium, guilty agents choose to confess the crime. However, these guilty agents are made indifferent between confessing and not to. So their punishment is indirectly determined by those second stage trial punishments. In that sense, it is as if every agent's punishment is determined by the second stage trial, which leads to the conclusion that, in general, the trial system is not optimal.

Proposition 3 *If $\phi = 1$, the trial system is second best optimal if and only if the agents' types are independent.*

The following example illustrates the insufficiencies of the trial system when the agents' types are not independent.

Example. *Suppose that $N = 2$ with a symmetric prior distribution of guilt being given by the following table:*

t_n t_{-n}	i	g
i	$\frac{1+\rho}{4}$	$\frac{1-\rho}{4}$
g	$\frac{1-\rho}{4}$	$\frac{1+\rho}{4}$

The parameter $\rho \in [-1, 1]$ determines whether there is negative or positive correlation between the agents' types. In particular, if $\rho < 0$ then $\pi\left(t_n = g|t_{-n} = i\right) > \pi\left(t_n = g|t_{-n} = g\right)$ and so there is negative correlation, while if $\rho > 0$ the opposite happens and there is positive correlation.

Assume further that $\pi\left(\theta_n|t_n = i\right) = 2\left(1 - \theta_n\right)$, $\pi\left(\theta_n|t_n = g\right) = 2\theta_n$ and $\alpha = 1$.

In the optimal trial system, any given agent n is punished in 1 if $\pi\left(t_n = g|\theta\right) > \frac{1}{2}$ and acquitted otherwise. It then follows that agent n is punished if and only if

$$\theta_n > \theta_n^{Tr}\left(\theta_{-n}\right) \equiv \frac{1}{2} + \rho\left(\frac{1}{2} - \theta_{-n}\right)$$

The impact that θ_{-n} has on θ_n^{Tr} depends very much on how correlated the agents' types are. Suppose that θ_{-n} is large. This means that it is likely that $t_{-n} = g$. If there is negative *correlation ($\rho < 0$) it follows that agent n is likely to be innocent and so θ_n^{Tr} is larger than $\frac{1}{2}$. If, on the contrary, there is positive correlation ($\rho > 0$) then agent n is more likely to be guilty and θ_n^{Tr} is smaller than $\frac{1}{2}$. This implies that*

$$B_n^i = \frac{1}{4} - \frac{1}{12}\rho^2$$

while

$$B_n^g = \frac{1}{12}\rho^2 + \frac{3}{4}$$

and so

$$V_n^{Tr} = \frac{1}{12}\rho^2 - \frac{1}{4}$$

The trial solution is better if there is more correlation because, in that case, θ_{-n} is more informative and enables the principal to make more accurate decisions.

Now consider the optimal CIS. If agent n decides to go to trial, the standard of proof will depend on the report of the other agent. In particular, it will be the case that if the other agent reports to be innocent, then agent n is punished if and only if

$$\theta_n > \theta_n^{SB}\left(t_{-n} = i\right) \equiv \frac{1+\rho}{2}$$

while if the other agent reports to be guilty, then agent n is punished if and only if

$$\theta_n > \theta_n^{SB}\left(t_{-n} = g\right) \equiv \frac{1-\rho}{2}$$

If the other agent reports to be innocent, then agent n is more likely to be guilty if there is negative correlation ($\rho < 0$). As a result, the standard of proof is reduced. If, on the contrary, there is positive correlation ($\rho > 0$), then the standard of proof is increased. The opposite happens when the other agent reports to be guilty. This implies that

$$B_n^i = \frac{1}{4} - \frac{1}{4}\rho^2$$

while

$$B_n^g = \frac{3}{4} + \frac{1}{4}\rho^2$$

and so

$$V_n^{SB} = \frac{1}{4}\rho^2 - \frac{1}{4}$$

Just like in the trial system, more correlation is beneficial for the principal because it allows her to select punishments that are more accurate. However, this benefit is magnified in the CIS because it is more effective in using the information provided by the other agent. In a CIS the second stage punishments are determined by the other agent's type while in the trial system they are determined by the other agent's evidence level. In particular, notice that as ρ converges to either 1 or -1 (as the correlation becomes perfect) the expected utility of the principal approaches the first best (V_n^{SB} converges to 0).

Figure 3 compares V_n^{Tr} and V_n^{SB} for different values of ρ.

There is one interesting property of any optimal allocation that I believe is worth emphasizing. Notice that condition (6) represents the optimal decision regarding agent n that the principal is able to make, given the information provided by all other agents and the evidence he is to collect. 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

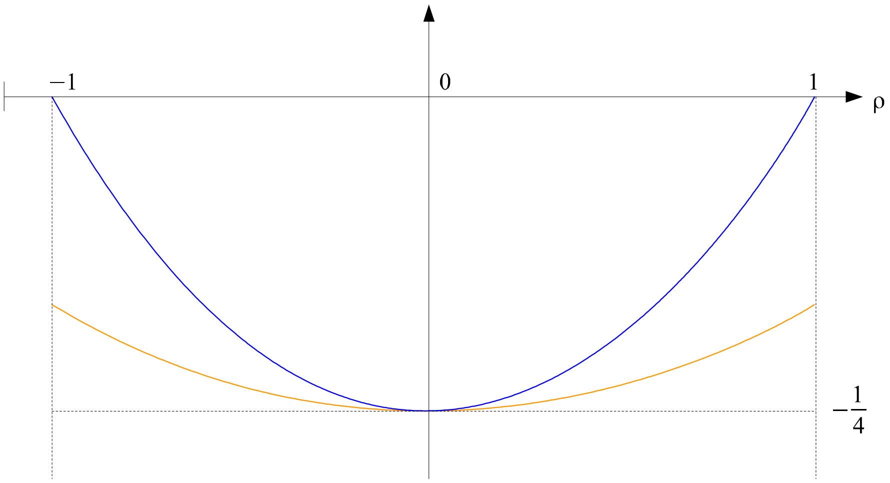

Figure 3: The orange and blue curves represent V_n^{Tr} and V_n^{SB} respectively, as a function of ρ

that he is guilty) will not be used against him when determining what punishment he is to get. So, in a way, the optimal allocation is in contrast with the American criminal law practice of the *Miranda warnings*, or at least 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 - he should be providing a guarantee that he will **not** use this information against the agent - which ironically enough, in the current legislation, is actually achieved by purposefully not reading the Miranda warnings. This feature is even more important when agents have additional information about the crime, like the identity of fellow criminals. I study this case in more detail in section 7.2.

For each α, let $x^{SB}(\alpha)$ denote the optimal second best allocation that is implemented by a CIS. In Figure 4, I display how the parameter α influences the expected punishment of any given agent under x^{SB}. Recall that α measures how important it is for the principal not to punish innocent agents, relative to her desire to punish guilty agents. If $\alpha = 0$, there is no concern with protecting innocent agents and, as a result, each agent is punished regardless of evidence. As α becomes larger, the expected punishment of innocent agents becomes smaller, which necessarily implies that the the expected punishments of guilty agents must also become smaller, for otherwise they would prefer to misreport. If α becomes large enough, the expected punishment of the agent converges to 0, regardless of whether he is innocent or guilty.

One of the differences from this paper to others in the Law and Economics literature is that I do not explicitly model the agents' decision of committing the crime.[7] In my analysis, I assume the crime has been committed already and the randomness of the agents' innocence (vector t) simply reflects the fact that the principal does not know the identity of the criminals. This description might leave the reader with the impression that my analysis does not consider the deterrence role that a criminal justice system is supposed to have. In particular, the assumed utility function of the principal does not directly take into account the concern the principal should have of deterring crime. Figure 4 is particularly useful in that it allows me to address these concerns in a clear way.

[7] See Garoupa (1997) for several of these examples.

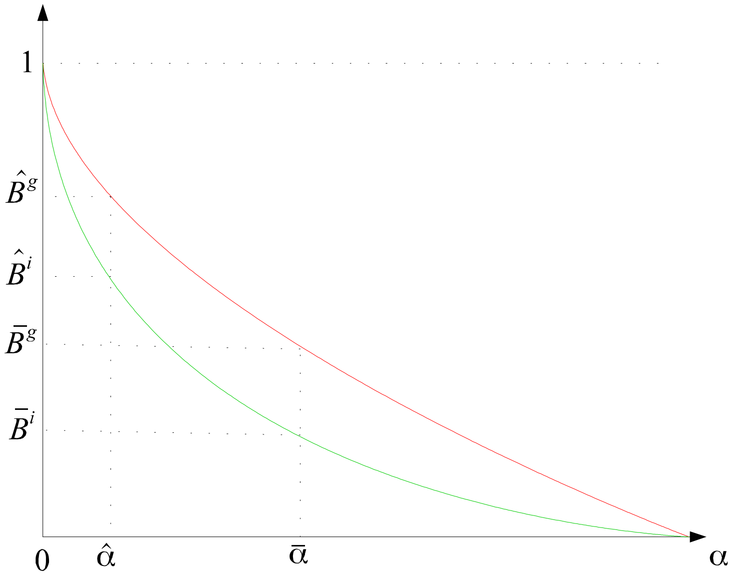

Figure 4: The green and red lines represent the expected punishment of a given agent when innocent and guilty respectively as a function of α.

Notice that Figure 4 identifies the set of "second-best efficient" points: for each level of expected punishment for the innocent agents (B^i), Figure 4 identifies the highest possible expected punishment the guilty agents might be given in any allocation (B^g). So, for example, if the principal's goal is to find an allocation that maximizes B^g subject to $B^i \leq \widehat{B}^i$, the answer is $x^{SB}(\widehat{\alpha})$, which results in $B^g = \widehat{B}^g$. This is because, if there was some other allocation with a higher B^g but the same B^i, that would be the optimal allocation under the preferences that I have assumed in this paper when $\alpha = \widehat{\alpha}$. Therefore, all preferences of the sort "maximize B^g subject to B^i" can be mapped into a given α and fall under my analysis. But now consider what the best way of deterring crime would be. If the principal wants to decrease the incentives to commit a crime, it should maximize the difference between the expected punishment that a guilty agent is to receive and that of an innocent - it should maximize $\left\{B^g - B^i\right\}$. It then follows, from Figure 4, that the allocation that maximizes $\left\{B^g - B^i\right\}$ is $x^{SB}(\overline{\alpha})$. Therefore, the case of a principal with deterrence concerns is a special case of my model, characterized by $\alpha = \overline{\alpha}$.[8]

The role of ϕ

Recall that the optimal punishment the principal wishes to impose on a guilty agent is 1. Therefore, the parameter ϕ can be interpreted as the ability the principal has to "overpunish" the agent. It is easy to see that this ability increases the expected utility of the principal, as a larger ϕ constrains the problem less.

Proposition 4 characterizes the optimal allocation x^{SB} for a general ϕ.

[8] Notice that it is possible that the value α that maximizes deterrence is not the same for all agents. But, that is resolved if one assumes that, for each n, there is a potentially different α_n. Given that the N problems are independent, all results are exactly the same.

Proposition 4 *For all n, for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$,*

$$\begin{cases} x_n^{SB}(i, t_{-n}, \theta) = \begin{cases} \phi \text{ if } \theta_n > \theta_n^{SB}(t_{-n}) \\ 0 \text{ otherwise} \end{cases} \\ x_n^{SB}(g, t_{-n}, \theta) = \phi \sum\limits_{t_{-n} \in T_{-n}} \frac{\pi(g, t_{-n})}{\pi(t_n = g)} \int\limits_{\theta_n^{SB}(t_{-n})}^{1} \pi(\theta_n | t_n = g) \, d\theta_n \end{cases}$$

where

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

and the constant $\lambda_n \in [0, 1)$ is completely characterized in the proof.

Proof. See appendix. ■

The type of solution is the same as with $\phi = 1$: all agents are given the opportunity to confess to have committed the crime in exchange for a constant punishment. Guilty agents choose to confess the crime, even though they are indifferent, while innocent agents prefer to proceed to the second stage, where they are to be punished if and only if the evidence level is sufficiently large.

There are three differences to the case of $\phi = 1$. First, if the agent is punished at the second stage trial, he is to receive a punishment of ϕ and not 1, i.e. he is to receive a punishment that is greater than the one that fits the crime. The intuition for this result is similar to that of the example of the Introduction. Because a guilty agent is relatively more affected by a reduction of the standard of proof than an innocent agent, it is always better to punish agents as harshly as possible at trial and then select the standard of proof (the threshold over the evidence level) as a function of the principal's preferences. The second difference has to with the threshold θ_n^{SB}. The constant λ_n is proportional to the lagrange multiplier associated with condition (5). Hence, if $\phi = 1$ then $\lambda_n = 0$. But if ϕ is sufficiently large (bigger than $\overline{\phi}_n > 1$ that is characterized in the proof of Proposition 4), then λ_n becomes positive and the threshold θ_n^{SB} becomes larger. Finally, the third difference is that if $\phi \geq \overline{\phi}_n$ then allocation x^{SB} is uniquely optimal.[9]

Figure 5 depicts the evolution of the solution as ϕ increases.

If ϕ is close to 1 - in Figure 5, if $\phi \leq \overline{\phi}_n$ - constraint (5) does not bind and $\lambda_n = 0$. Therefore, the standard of proof used at the second stage trial is equal to the one when $\phi = 1$. This means that increases of ϕ do not impact the likelihood an agent is punished at trial but increase the punishment itself, in the event of a conviction. Hence, the innocent's expected punishment is increased, because he chooses to go to trial, and the guilty's expected punishment is also increased, because, even though he does not go to trial, he is made indifferent. So, if $\phi \leq \overline{\phi}_n$, the expected punishment of the agent increases, regardless of whether he is innocent or guilty. As ϕ increases, the expected punishment of the agent when guilty reaches 1, which happens at $\phi = \overline{\phi}_n$. For $\phi > \overline{\phi}_n$, the constraint begins to bind. Because the expected punishment of the agent must be 1 when he is guilty, and the punishment at trial is growing with ϕ, it must be that the probability of conviction at trial is becoming smaller - so λ_n is strictly increasing for all $\phi \geq \overline{\phi}_n$. So much so that the innocent's expected punishment is becoming smaller - recall the example in the Introduction where it was possible to decrease the expected punishment of the agent when innocent while keeping it constant when guilty, by continuously increasing the second stage punishments. Proposition 5 shows that, for all N, this process of increasing ϕ leads to the first best solution.

Let $B_n^{t_n}\left(x_n^{SB}\right)$ denote the expected punishment of agent n when his type is t_n under allocation x^{SB}.

[9] Recall that the simplified problem does not depend on guilty punishments. The only requirement is that the expected punishment for the guilty agent to be equal to B_n^g. If it is optimal to set $B_n^g = 1$, the only way this happens is if all punishments are equal to 1, because it is not optimal to punish guilty agents in more than 1 in any event.

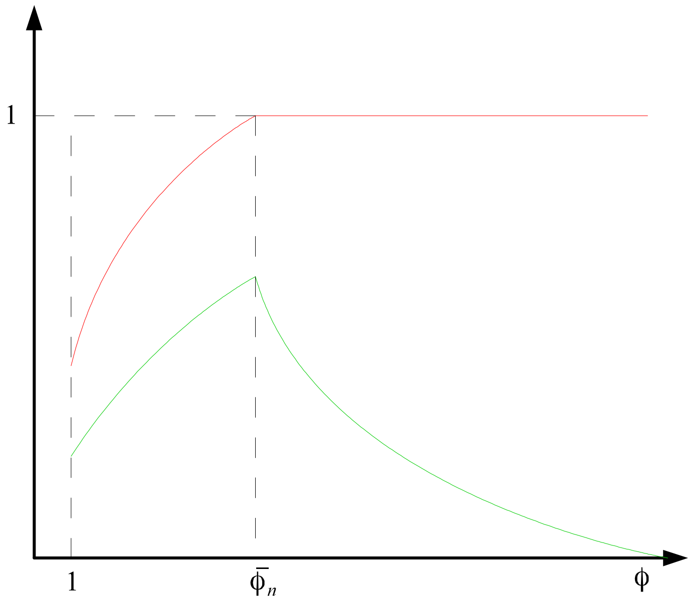

Figure 5: Evolution of the agent's expected punishment as a function of ϕ. The red and green curves represent the expected punishment when the agent is guilty and innocent respectively.

Proposition 5 *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. ■

Proposition 5 states that, as long ϕ is sufficiently large, it is possible to build an incentive compatible mechanism that approximates the first best allocation. This result is reminiscent of Cremer and McLean (1988), where it is shown that, if an agent's type affects his beliefs about other agent's types, then, under some conditions, it is possible to implement the principal's preferred outcome. In Cremer and McLean (1988), an agent's type affects his beliefs because agents' types are not independent. In this paper, however, even if agents' types are independent, Proposition 5 holds. The reason is that the analysis also includes evidence. Innocent and guilty agents have different beliefs with respect to what evidence is likely to be generated, and a similar type of argument to the one of Cremer and McLean (1988) can still be used. The idea is that, if ϕ is very large, the principal can simply punish in 1 all guilty reports, and set a lottery of punishments, as a function of all agents' reports and evidence, that simultaneously gives an expected punishment close to 0 to innocent agents that report truthfully and an expected punishment of 1 to guilty agents that choose to misreport. The principal is able to do this by, at the second stage trial, punishing agents only if the evidence level is very close to 0, so that it is infinitely more likely for it to have been generated by a guilty agent.

The problem of excessive commitment power

The CIS that implements x^{SB} is based on the assumption that the principal is able to commit to a set of allocations, even after observing agents' reports and evidence. That assumption allows the principal i) not to punish guilty agents in 1 once they confess, and ii) to punish innocent agents even with the knowledge they are indeed innocent.

As for i), only guilty agents confess the crime in equilibrium. Hence, upon hearing a confession, the principal would prefer to renege his promise and punish the agent in 1. Of course, knowing this, a guilty agent would not confess. Is it reasonable to believe 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.[10] 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, all innocent agents choose not to confess to have 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 knowingly innocent agents. This is harder to accept as, not only does the principal prefer to go back on his promise of punishment, but also the agent prefers he 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 hopes that the promise of punishment would be reneged by the principal. Even if the principal employed such a system through law it is still unlikely that a society is willing to accept that knowingly innocent agents are to be punished, particularly given the human element that is present in the appreciation of the evidence.

In the next section, I address the same problem but assume the principal has limited commitment power. I analyze the problem of constructing an optimal criminal justice system under two different assumptions. First, I analyze renegotiation proof mechanisms - mechanisms that principal and agents do not wish to renegotiate - which eliminates implication ii) - knowingly innocent agents are no longer punished, for otherwise they would have rather renegotiate and eliminate such punishment. Second, I analyze sequentially optimal mechanisms, where the principal has no commitment power and is free to decide punishments without being restricted by any promise, which not only eliminates implication ii) but also implication i) - knowingly guilty agents are punished in no less than 1.

6 Limited Commitment Power

In this section, I analyze the problem the principal faces of constructing a criminal justice system when he has limited commitment power. I first analyze renegotiation proof mechanisms and then sequentially optimal mechanisms. In either case, because the principal has limited commitment power, the revelation principle no longer holds, which means that, in general, it is not enough to consider only revelation mechanisms.

The timing is as in the previous section. Before any evidence is generated the principal selects a mechanism. Given the mechanism, each agent n simultaneously chooses to send his preferred message from the message set M_n, prior to knowing the evidence. Let $M = M_1 \times ... \times M_N$ and refer to m_n as a generic element of M_n and m of M. I also give the usual interpretation to $m_{-n} = (m_1, ..., m_{n-1}, m_{n+1}, ..., m_N)$ and $M_{-n} = M_1 \times ... \times M_{n-1} \times M_{n+1} \times ... \times M_N$. A mechanism $x : M \times \Theta \rightarrow \mathbb{R}_+^N$ is a map from the agents' messages and from the evidence to punishments.

Given the mechanism, each agent selects a probability distribution over his message space for each type. Agent n's strategy is denoted by $\sigma_n : \{i, g\} \times M_n \rightarrow \mathbb{R}_+$ and the strategy profile by $\sigma = (\sigma_1, ..., \sigma_N)$. Finally, the set of all of strategy profiles is denoted by Φ.

I call each profile (x, σ) a system and evaluate it through the principal's expected utility. I

[10] 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.

denote by $\widehat{V}(x,\sigma)$ the principal's expected utility of pair (x,σ), where

$$\widehat{V}(x,\sigma)=\sum_{t\in T}\int_{\theta\in\Theta}\int_{m\in M}\pi(t,\theta)\,\sigma(t,m)\,u^{p}(t,x)\,dmd\theta$$

Strategy profile $\sigma\in\Phi$ 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

$$\begin{aligned}&-\int_{\theta\in\Theta}\int_{m_{-n}\in M_{-n}}\pi^{\sigma}(m_{-n},\theta|t_n)\,x_n(m_n,m_{-n},\theta)\,dm_{-n}d\theta \\ \geq\ &-\int_{\theta\in\Theta}\int_{m_{-n}\in M_{-n}}\pi^{\sigma}(m_{-n},\theta|t_n)\,x_n(m_n',m_{-n},\theta)\,dm_{-n}d\theta \text{ for all } m_n'\in M_n\end{aligned}\tag{7}$$

where $\pi^{\sigma}(m_{-n},\theta|t_n)$ represents the conditional joint density of (m_{-n},θ), given agent n's type t_n and strategy profile σ. If condition (7) holds, I say that the system (x,σ) is incentive compatible.

It is also convenient to formally define a concept I have used throughout, in light of the notation presented.

Definition 6 *A CIS (x,σ) is such that, for all n,*

i) Only two messages are sent in equilibrium by each agent: a confessing message c and a non-confessing message.

ii) If an agent confesses, he receives a constant punishment - $x_n(c,m_{-n},\theta)$ is independent of all $m_{-n}\in M_{-n}$ and $\theta\in\Theta$.

Finally, if (x,σ) constitutes a CIS, I refer to x as a confession inducing mechanism.

6.1 Renegotiation Proof Mechanisms

What defines a renegotiation proof mechanism is that, after observing any (m,θ), the principal is unable to reach an agreement with any agent to alter the promised punishment in a way that is mutually beneficial. Consider system (x,σ). Given strategy profile σ and after observing (m,θ), the principal will form a belief about agent n's type, given by $\pi^{\sigma}(t_n|m,\theta)$. Let $\gamma_n^{\sigma}(m,\theta)$ be the optimal punishment the principal would like to impose on agent n, given such beliefs, i.e.[11]

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

If $x_n(m,\theta)>\gamma_n^{\sigma}(m,\theta)$ - if the punishment imposed on agent n is higher 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 lower punishment.

Definition 7 *The system (x,σ) is renegotiation proof if and only if, for all n,m and θ,*

$$x_n(m,\theta)\leq\gamma_n^{\sigma}(m,\theta)\tag{8}$$

[11] If there are multiple maximizers, $\gamma_n^{\sigma}(m,\theta)$ takes the smallest one.

If system (x,σ) is renegotiation proof, then I say that mechanism x is renegotiation proof. Notice that the CIS described in the previous section that implements x^{SB} is not renegotiation proof. The strategy profile considered 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.

I start the analysis of the optimal renegotiation proof system by stating Lemma 8 that delimits the message set of each agent.

Lemma 8 *Without loss of generality, it is possible to set $M_n = \mathbb{R}_+ \cup \{c\}$ for all n.*

Proof. See appendix. ■

The meaning of a message is given by the belief the principal forms when she receives it. In Lemma 8, I show that any two given messages that generate the same posterior belief can be reduced to a single one. In particular, if, for any given agent n, there are two messages m_n' and m_n'' such that $r_n(m_n') \equiv \frac{\sigma_n(g,m_n')}{\sigma_n(i,m_n')} = \frac{\sigma_n(g,m_n'')}{\sigma_n(i,m_n'')} \equiv r_n(m_n'')$, then it is possible to construct an equivalent system with only one of those two messages. Hence, 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 characterize the optimal renegotiation proof system $\left(x^{RP}, \sigma^{RP}\right)$ in two steps. First, in Lemma 9, for all σ, I characterize the optimal allocation x^σ so that $\widehat{V}(x^\sigma,\sigma) \geq \widehat{V}(x,\sigma)$ for all x such that (x,σ) is incentive compatible and renegotiation proof. Then, in the second step, in Proposition 10, I characterize σ^{RP} and show that $\left(x^{RP}, \sigma^{RP}\right)$ constitutes a CIS.

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 9 *For all n,*

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

where

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

Proof. See appendix. ■

One can think of x^σ as a two stage mechanism, where, in the first stage, agents are given the opportunity to confess (send message c) or not (and send one of the other messages). If agent n confesses, he receives a constant punishment of φ_n. If he does not confess, then his punishments are determined in the second stage. In that case, if the agent has sent message m_n^σ, the principal is supposed to choose his preferred punishment conditional on what he has learned in the first stage and on the evidence - $x_n^\sigma(m_n^\sigma, m_{-n}, \theta) = \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta)$. If the principal was to do the same when the agent sends other messages, these punishments would be larger than those after m_n^σ, which would not be incentive compatible. Hence, for these messages, the principal chooses punishments that are close as optimal as possible - $x_n^\sigma(m_n, m_{-n}, \theta) = \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta)$ for all $m_n \in \mathbb{R}_+$.

Notice that a CIS is a simplified version of this mechanism in that there is only one non-confessing message sent by each agent.

Proposition 10 *A CIS is optimal within the set of incentive compatible and renegotiation proof systems.*

Proof. See appendix. ■

In proposition 10 I show that it is optimal for agents to send at most two messages: the confessing message c and a non confessing message $\bar{c}$. The argument is as follows. Take any system σ and label message m_n^σ as $\bar{c}$. Suppose that, without loss of generality, agent 1 is sending a second non-confessing message m_1' in addition to message $\bar{c}$. As mentioned above, a message is identified by its "guiltiness" ratio $\frac{\sigma_1(g,m_1)}{\sigma_1(i,m_1)} \equiv r_1(m_1)$. Suppose that $r_1(\bar{c}) < r_1(m_1') < \infty$. The idea of Proposition 10 is that by shifting weight v from $\sigma_1(g, m_1')$ to $\sigma_1(g, c)$ enough so that $\frac{\sigma_1(g,m_1')-v}{\sigma_1(i,m_1')} = r_1(\bar{c})$, it is possible to increase the expected utility of the principal (see Figure 6).

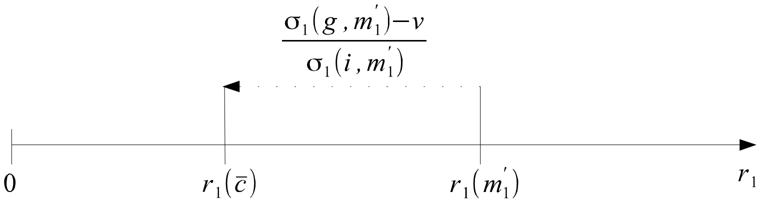

Figure 6: Shift from $r_1(m_1')$ to $r_1(\bar{c})$

The expected punishment of agent 1 is unchanged regardless of whether he is innocent or guilty, because message $\bar{c}$ is still available and the expected punishment of sending it remains the same (ratio $r_1(\bar{c})$ is unchanged). The difference, though, is that the expected utility the principal is able to retrieve from any of the other agents is now increased. The logic is similar to the previous section. In the event that agent 1 is guilty, by confessing more often, he makes it more likely that the principal has more accurate information when choosing the other agents' punishments.

The conclusion of Proposition 10 is that a CIS is still optimal even when the principal has reduced commitment power. It is a different CIS than the one of the previous section in that the second stage punishments are sequentially optimal. In the previous section, the second stage punishments were chosen regardless of the perceived guilt of the agent. In particular, when agents report truthfully and innocent agents refused to confess, the principal was still supposed to punish them in the second stage. He was only able do this because he was able to commit, which would mean having a set of laws and regulations for judges, lawyers and jurors to follow that are not necessarily designed to assess the agents' guilt. But under this new CIS this is no longer necessary. Implementing such a CIS requires only the guarantee that the rights of confessing agents are protected.

Finally, notice that a trial system can be thought of as a CIS in which no agent chooses to confess. In Proposition 11 I show that such a system is not optimal.

Proposition 11 *The trial system is* ***not*** *an optimal renegotiation proof system unless agents' have independent types.*

Proof. See appendix. ■

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 is able to choose more appropriate punishments. There is also an indirect impact in that the beliefs of the principal are now slightly altered in the event that agent 1 does not confess, which might decrease the expected utility the principal retrieves from agent 1. Proposition 11 shows that, if the probability of confession is sufficiently small, it is possible to guarantee that the direct impact dominates. I end this section by continuing the example of the previous section.

Example (continued) *Assume now that the principal has limited commitment power and is no longer able to commit not to renegotiate. In the optimal CIS that implements x^{RP} innocent agents do not confess while guilty agents confess with probability $z_n \in [0,1]$. Consider the punishments of agents that choose not to confess. If the other agent does not confess the crime (chooses to play $\bar{c}$), agent n is punished if and only if*

$$\theta_n > \theta_n^{RP}(\bar{c}, \theta_{-n}) = \frac{(1-\rho)(1-z_{-n})\theta_{-n} + (1+\rho)(1-\theta_{-n})}{\left[\begin{array}{c}(1+\rho)(1-z_n)(1-z_{-n})\theta_{-n} + (1-\rho)(1-z_n)(1-\theta_{-n}) \\ +(1-\rho)(1-z_{-n})\theta_{-n} + (1+\rho)(1-\theta_{-n})\end{array}\right]}$$

while if the other agents chooses to confess (chooses to play c), then agent n is punished if and only if

$$\theta_n > \theta_n^{RP}(c) \equiv \frac{1-\rho}{(1+\rho)(1-z_n) + 1 - \rho}$$

Notice that if $z_1 = z_2 = 0$ this CIS becomes the trial system in that no agent confesses and threshold $\theta_n^{RP}(\bar{c}, \theta_{-n})$ becomes equal to $\theta_n^{Tr}(\theta_{-n})$. As for the connection with the second best allocation it follows that the first threshold is only equal to $\theta_n^{SB}(t_{-n} = i)$ if $z_n = 0$ and $z_{-n} = 1$. This means that either $\theta_1^{RP}(\bar{c}, \theta_{-n}) \neq \theta_1^{SB}(t_{-n} = i)$ or $\theta_2^{RP}(\bar{c}, \theta_{-n}) \neq \theta_2^{SB}(t_{-n} = i)$ if $\rho \neq 0$. It then follows that, unless there is no correlation between the agents' types, the principal is strictly worse off by having reduced commitment power. Figure 7 adds the expected utility the principal gets from the optimal renegotiation proof allocation x^{RP} (denoted by V^{RP}) to Figure 3.

Once again, more correlation between the agents' types, being it positive or negative, increases the expected utility of the principal because it makes the information each agent provides more important, which means that there are larger information externalities in each confession.

6.2 Sequentially Optimal Mechanisms

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

Recall that $\gamma_n^\sigma(m, \theta)$ denotes the optimal punishment the principal would like to impose on agent n, given strategy profile σ, and after observing message m and evidence θ. If the principal has no commitment power, he must act optimally for every (m, θ) he observes.

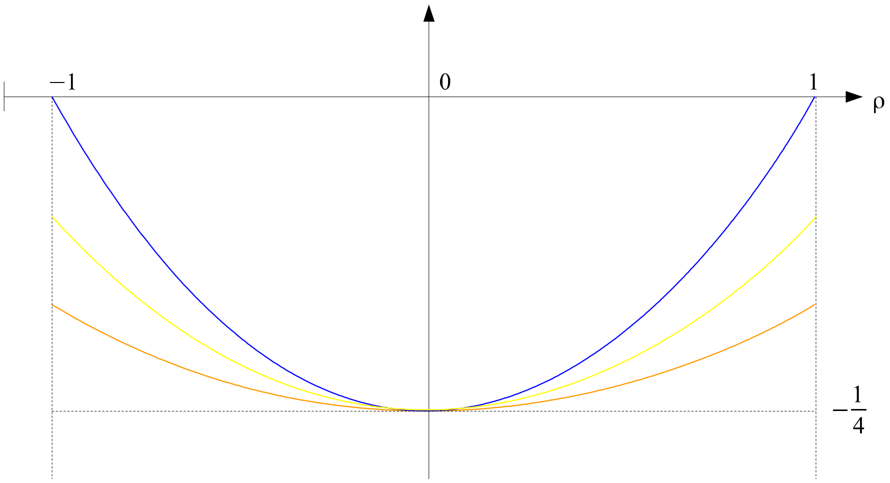

Figure 7: The orange, yellow and blue curves represent V_n^{Tr}, V_n^{RP} and V_n^{SB} respectively, as a function of ρ

Definition 12 *The system (x, σ) is sequentially optimal if and only if, for all n, m and θ,*

$$x_n(m, \theta) = \gamma_n^{\sigma}(m, \theta)$$

By eliminating the commitment power of the principal, one also eliminates his 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 13 *If the principal has no commitment power, a trial system is optimal.*

6.3 How much commitment power does the principal have?

This paper characterizes the principal's preferred mechanism under three different assumptions regarding his 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 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 that is granted the exclusive responsibility of administering criminal justice and make him the principal in the model. But if the benevolent dictator is the principal, he 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 him. So he 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 case see that there are several examples where leniency is exerted towards agents that confess to have committed a crime: for example, plea bargaining is a common practice in the United States. The method modern societies seem to follow, in order to commit to exert 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 the prosecutor cannot go back on his word once he has obtained the confession from the agent. But if societies can use law to create commitment power, 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 allocation under full commitment power. This allocation requires that innocent agents are to be punished if their evidence level is too low. By the nature of the mechanism that implements it, it is known that the agents 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 the agent is innocent (from observing he chose not to confess to have committed the crime), it seems reasonable to believe they would always claim the evidence level is low to avoid convicting him.

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 has not testified. As Laudan (2006) points out, this practice precludes important information from the trial and seems to be an attempt at conditioning 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 and the optimal allocation given by x^{SB}. If not, then one accepts the principal has some limited commitment power and is only able to implement x^{RP}. Recall that both systems involve two stages: a first stage where agents may choose to confess and receive an immediate punishment, followed by a trial of the non-confessing agents. The key difference is precisely that only under x^{RP} is the assessment of guilt "honest" at trial, in that agents are judged using all information available at the time.

7 Extensions

The main purpose of this paper is to highlight the virtues of CISs in particular when the agents' guilt is correlated. In the main text, I have presented the simplest possible model that made my argument clear. There were, however, several simplifications that might leave the reader wondering about the robustness of the results. In this section, I extend the original model and the analysis of section 5 on the second best problem in order to address some of these concerns.

I divide this section into four parts. In the first extension, I allow the agents and the principal to be risk averse. In this case, one might think that CISs might no longer be appealing because it might be the case that agents confess not because they are guilty but because they are risk averse. I show that this is not the case if the principal is aware of how risk averse the agents are. In fact, I show that, in this case, not only are CISs still optimal but they are actually uniquely optimal.

In the second extension, I consider a more general information structure where each agent might be a part of a conspiracy to commit the crime, and so be informed about the identity of the other conspirators. In this case, the correlation between agents' types is even more evident, which makes it more clear that the trial system is not optimal. I show that, in this framework, the optimal system is an *extended* CIS in which agents that confess are also requested to report what they know about the crime, without having that information being used against them when being punished.

As discussed in section 5, one of the issues of the optimal CIS when the principal has commitment power is that there is a perfect separation between those that are guilty, who choose to confess, and those that are innocent, who choose not to. In section 6, by limiting the commitment power of the principal, I have shown that such feature disappears and that both guilty and innocent agents might refuse to confess. In the third extension, I argue that, even if one still assumes the principal has commitment power, in general, it is not the case that there is perfect separation between those that are guilty and those that are innocent. In particular, I argue that if one allows for privately observed heterogeneity in the way agents perceive the evidence, it is either not possible or not desirable for the principal to design a CIS that guarantees that all guilty agents confess and that all innocent agents do not.

Finally, in the fourth extension I consider a change in the timing of the mechanism selection by the principal. Rather than being able to select a mechanism before knowing the evidence, I consider the case where she can only do so after having observed it. This particular problem is usually referred to in the literature as an informed principal's problem[12].

7.1 Risk Averse Agents

One of the assumption of this paper is that agents are risk neutral. This might lead the reader to inquire on whether CISs 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.

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 $u^i(x_n) = -x_n^{\omega_i}$, $u^g(x_n) = -x_n^{\omega_g}$ 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 allocation that can be implemented by a trial system.

Proposition 14 *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. ■

In the 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.

[12] The classic references on the informed principal literature are Myerson (1983) and Maskin and Tirole (1990).

Let $\widetilde{x}^{SB}$ denote the second best allocation - optimal within the set of incentive compatible allocations.

Proposition 15 *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. ■

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, he 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 he knows him to be innocent. This assumption is convenient in that it guarantees that innocents' incentive constraints do not bind.

Proposition 15 implies that, if the agents and the principal are risk averse, the optimal allocation is implemented by 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 allocation, 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 punishment 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 allocation, a guilty agent receives a lottery of distinct punishments. If the guilty agent is strictly risk averse he would be willing to accept a constant punishment that is larger than the expected punishment of the original lottery. The principal would prefer this alternative if she is risk averse or even if she is risk neutral. So, a sufficient condition for this result is that guilty agents are strictly risk averse. But even if guilty agents are risk neutral this change is still beneficial for the principal if he is strictly risk averse. By guaranteeing that guilty agents receive a constant punishment, the principal reduces the risk of letting guilty agents escape.

Notice that, if agents and principal are risk averse, the case for CISs 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 CISs 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 he is able to select punishment allocations 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 allocation depends on the risk aversion level of innocent and guilty agents.

Proposition 16 *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

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

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

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

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. ■

When the principal is determining the optimal punishments to impose on innocent agents he faces a trade-off. On the one hand, he 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 that are innocent but very unappealing for those that are guilty. If innocent agents are more risk averse than guilty agents, then smoothing punishments is relatively better for innocent agents rather than guilty ones, 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 than to innocent ones. 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.

7.2 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 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) has been realized. Agent 3 is innocent and so 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. 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 the principal has commitment power.

Let $L \subset \widehat{T}$ 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 allocation $\widehat{x}^{SB}:\widehat{T}\times\Theta\rightarrow\mathbb{R}_{+}^{N}$, where $\widehat{T}=\widehat{T}_1\times...\times\widehat{T}_N$, be defined as follows. For all $\widehat{t}\in L$, $\theta\in\Theta$ and for all n, $\widehat{x}_n^{SB}(t,\theta)=1$.

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,

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

where φ_n is characterized in the proof of Proposition 17.

Proposition 17 *Allocation $\widehat{x}^{SB}$ is optimal within the set of incentive compatible allocations.*

Proof. See appendix. ■

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 his best interest to punish the agents as much as possible. The rest of the allocation 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. The allocation is implemented by an *extended* CIS. In the first stage, and in the same way as 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 or not, while innocent agents prefer not to. These proceed to the second stage and are judged only with the information the principal can gather from other agents. Another feature of this system is that agents that 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 that 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.

I illustrate how this extended CIS works by continuing the example of sections 5 and 6.

Example (continued) *Consider the case where $p([i,g])=p([g,i])=0$ and $p([g,g])=\varsigma\in(0,1)$ and, for ease of exposition, assume $\rho=-\frac{1}{2}$. One can think of this scenario as representing the fire example in the Introduction when the principal has 2 suspects. One possibility is that only one of the agents committed the crime - $t=(i,g)$ or $t=(g,i)$. In this case, it is assumed that the guilty agent does not know whether the other agent is also guilty. The logic of this assumption is that if an agent individually decides to start the fire he does not have a conspirator and so has no way of knowing whether, in some other location of the forest, the other agent is also starting a fire by himself. If both agents confess the crime, while it is certainly possible that both agents act independently, it is also likely that they conspire to commit the crime. So, the assumption is that there would be a probability of ς of the latter scenario occurring. Notice that if $\varsigma=0$ we are back to the previous section where each agent knows only their type.*

Without loss of generality take the case of agent 1. If agent 2 incriminates him (reports he is of type (g,g)) agent 1 is bound to receive a punishment of 1. If he reports truthfully, the principal knows that the report of agent 2 is valid and punishes agent 1 in 1. If he chooses to lie, then the principal becomes aware that one of the two agents is not reporting truthfully and punishes them both in 1. If agent 2 does not incriminate him, then, if agent 1 chooses to go *to trial, he is punished only if the evidence is sufficiently incriminatory. In particular, if agent* 2 *reports* $\widehat{g}$, *agent* 1 *is punished if and only if*

$$\theta_1 > \widehat{\theta}_1^{SB}\left(\widehat{t}_{-1} = \widehat{g}\right) \equiv \frac{3}{4-\varsigma}$$

while if agent 2 *reports* i, *agent* 1 *is punished if and only if*

$$\theta_1 > \widehat{\theta}_1^{SB}\left(\widehat{t}_{-1} = i\right) \equiv \frac{1}{4}$$

This leads to

$$\widehat{V}_1^{SB} = \frac{1}{8}\varsigma + \left(\frac{45}{16\varsigma - 64} - \frac{1}{2}\frac{\varsigma - 1}{(\varsigma-4)^2\left(\frac{1}{2}\varsigma - 2\right)}\left(\varsigma^2 - 8\varsigma + 7\right)\right)\left(\frac{1}{8}\varsigma - \frac{1}{2}\right) - \frac{3}{32}\frac{(\varsigma-1)^2}{\left(\frac{1}{2}\varsigma - 2\right)^2} - \frac{73}{128}$$

which is strictly increasing with ς. *Notice that* $\lim_{\varsigma\to 0}\widehat{V}_1^{SB} = V_1^{SB}$.

There are a few commentaries in order. First, the fact that this extended CIS takes into account that agents might have more information about the crime than merely whether they are guilty makes it preferred to the standard CIS, because it allows the principal to select more accurate punishments. As the example illustrates the more likely it is that agents know the identity of their co-conspirators (the larger is ς) the more likely it is they end up incriminating them, which is beneficial for the principal.

Second, all else the same, agents that commit a crime individually receive a lower expected punishment than those who belong to a criminal group. Of course there are other advantages to being part of a criminal organization - like benefitting from economies of scale - so this is not to say that organized crime is inefficient when looked at from the eyes of a criminal. It is rather to point out that my model's conclusions are very much in line with the intuition that agents who have committed a crime as part of a criminal group face additional risk: that the other criminals incriminate them. The fact that the agents themselves are aware of such risk only builds on the fear that someone else will confess (an agent who knows his fellow criminal is thinking about confessing is likely to confess himself), which is what makes the principal successful.

The third aspect that I believe is interesting is that members of a conspiracy are always punished in 1 because they are always incriminated. Remember that the idea of this mechanism is that the punishments that agents receive depend only on what other agents report (in addition to their own evidence level). It then follows that any agent that is a part of a conspiracy not only incriminates all other members but is also incriminated by them.

One problem with this argument though is the presence of multiple equilibria. In particular, in the case of the example, when both agents commit the crime together and know each other's identity they would both be better off if they both simultaneously deviated and reported to be innocent. This possibility of joint deviation seems even more plausible if we think the deviating agents must have been in contact in order to commit the crime together in the first place. However, it is easy to slightly alter the mechanism in order to eliminate this alternative equilibrium without decreasing the expected utility of the principal. I illustrate by continuing the previous example.

Example (continued) *Suppose that, in the event that both agents are guilty of committing the crime and know the other agent is also guilty* ($\widehat{t}_n = (g, g)$ *for* $n = 1, 2$), *agent* 2 *decides not to confess. Under allocation* $\widehat{x}^{SB}$ *agent* 1 *would no longer wish to report to be of extended type* (g, g) *as that would be understood as a lie* ($\widehat{t} = ((g, g), i) \in L$) *and would lead to a punishment of* 1. *In fact, agent* 1 *would have enough incentives to report to be innocent. In order for him* *not to, it is necessary to reward him by granting him a smaller punishment for confessing and incriminating agent 2 when agent 2 claims to be innocent . However, if one lowers agent 1's punishment unconditionally then he would incriminate agent 2 even when he does not know agent 2 is guilty. Hence, this reward should only be granted if the evidence of agent 2 supports agent 1's claim. In particular, let*

$$x_1\left(\left(g,g\right),i,\theta\right)=\left\{\begin{array}{l}1 \text{ if } \theta_2<d_1\\ 0 \text{ otherwise}\end{array}\right.$$

where $d_1 \in (0,1)$. Notice that if agent 1 does not know whether agent 2 is guilty it will be less appealing to report (g,g) when agent 2 reports innocent. Therefore, it is possible to select d_1 to guarantee that only when agent 1 knows agent 2 to be guilty does he choose to incriminate him. In particular, given the structure of the example, $d_1 \in \left(\frac{16-4\varsigma-\sqrt{\varsigma^2-56\varsigma+64}}{16-4\varsigma},\frac{\sqrt{15}}{4}\right)$. In this way, the truth telling equilibrium still exists and all punishments that occur with positive probability in that equilibrium remain unchanged, which means that the principal's expected payoff remains the same.

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. This makes the mechanism more robust and more likely to effectively punish conspiracy members.

7.3 Heterogeneous agents

In the model, 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 leave incriminating evidence. It can also be that agents are unlucky and leave some evidence behind - maybe someone who robbed a bank dropped their wallet in the escape. Even for innocent agents, it is likely that they too 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 his fingerprints will be found. One way to extend the model to allow for this type of agent heterogeneity is to assume that each agent n is privately informed of a continuous random variable $\beta_n \in [0,1]$ that determines the distribution of the evidence. In particular, let

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

denote the conditional distribution of θ_n given that the agent's β_n. The guilt (or innocence) of each agent influences θ_n only indirectly through β_n. In particular, assume that $\frac{\pi(\beta_n|t_n=g)}{\pi(\beta_n|t_n=i)}$ is strictly increasing for all $\beta_n \in [0,1]$, so that larger values of β_n are more likely if the agent is guilty than if he is innocent and so guilty agents are more likely to be the ones to observe incriminating evidence.

Proposition 18 below provides a characterization of the optimal CIS in this framework when $\phi = 1$ and the principal has commitment power.

Proposition 18 *For all n, there is $\left(\beta_n^i, \beta_n^g\right) \in [0,1]^2$ such that for all $t_n \in \{i, g\}$ and $\beta_n \in [0,1]$,*

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

where $s_n\left(t_n, \beta_n\right) \in \{c, \overline{c}\}$ represents the action that agent of type t_n with β_n chooses.

Proof. See appendix. ■

Agents that have a larger β_n are more likely to generate more incriminating evidence. Hence, they have a bigger incentive to confess (and select action c) than those with a lower β_n. If the agents' types are 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 must be 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' types (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 that confess (the following argument could also be made if only a small fraction of innocent agents confess). 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. Hence, 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, but now agents that confess must also face the same lottery of punishments they would have if they chose not to confess. They still have enough incentives to confess (because they are indifferent) but now their expected punishment is 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 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.

7.4 Informed Principal

I consider the same setup as in section 5 but now suppose the principal selects the mechanism *after* having observed evidence θ, which becomes his own private information. In this case, and based on the revelation principle, given each θ, the principal selects a mechanism $y_\theta : T \times \Theta \rightarrow$

$[0,1]^N$ that maps the agents' types and evidence to punishments. 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 the 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. For simplicity, I assume that $\phi = 1$.

Notice that any y that is a part of a PBE implements an allocation $x_y : T \times \Theta$ where $x_y(t,\theta) = y_\theta(t,\theta)$. I say that allocation x is incentive compatible when the principal acts *after* observing the evidence if there is a y that is part of a PBE such that $x = x_y$.

Proposition 19 *Any allocation $x : T \times \Theta$ that is incentive compatible when the principal acts after having observed the evidence is also incentive compatible when he acts before having observed the evidence.*

Proof. See appendix. ■

The intuition for this result is as follows. If x_y is incentive compatible when the principal acts after having observed the evidence then y is a part of a PBE. This implies that after each θ, when the principal selects mechanism y_θ, all agents prefer to tell the truth than not to. But if that is the case, then it must be that the expected utility of telling the truth is also bigger than not to, where the expectation is taken with respect to the realized θ. Hence, the original set of incentive constraints (IC) would necessarily be satisfied.

The implication of proposition 19 is that, if the principal is able to, she should act before she observes θ (or before θ is realized) and commit not to alter the mechanism upon observing it.

The opposite statement is not true. There are allocations that are incentive compatible when the principal acts before the evidence has been realized that would not be incentive compatible if he had acted afterwards. One such example is x^{SB}. Recall that x^{SB} specifies a constant punishment for the guilty agent, independent of evidence and other agents' reports. Suppose the principal chooses to act after having observed the evidence and that, for some n, the realized θ_n happens to be very small. In that case, the principal will be convinced that agent n is guilty with a high probability and so, it will be in his best interest to punish him in more than what is specified by x^{SB}.

Even though x^{SB} is not implementable if the principal acts after having observed the evidence, it is still possible for the principal to implement somewhat appealing allocations. Consider an allocation x^{IP} where, for all n, $x_n^{IP}(t_n, t_{-n}, \theta) = x_n^{SB}(i, t_{-n}, \theta)$ for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$.

Proposition 20 *x^{IP} is incentive compatible when the principal acts after the evidence.*

Proof. See appendix. ■

Recall that, from section 5, when $\phi = 1$, allocation x^{IP} is second best optimal as the punishments imposed on innocent agents are optimal by definition and the expected punishments of guilty agents make them indifferent to misreporting. The principal is able to implement this allocation by proposing it regardless of the θ she observes. In particular, her strategy is given by y^{IP} where $y^{IP}_{\widehat{\theta}}(t,\theta) = x^{IP}(t,\theta)$ for all t and θ, and for all $\widehat{\theta}$. This result implies that it is still possible for the principal to attain the same expected utility as in the second best solution, even though CISs are no longer optimal.

8 Concluding remarks

The main purpose of this paper is to argue for the virtues of CISs. The idea is that there are information externalities generated by 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. In fact, the sooner in the criminal process people are able to confess, the better, as, at that point, a confession is more valuable (given that less is known about the crime) and it is easier to induce guilty agents to confess (as they are more afraid that a future investigation might incriminate them). Even though this might appear as a radical suggestion, there are variants of CISs 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 that are able to confess to have broken an environmental regulation in exchange for a smaller punishment. 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 confessing, according to my analysis. A confession produces no externalities if there are no other agents to consider. In that sense, my analysis can be used as an argument to have plea discussions earlier in the criminal process, 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 that are not directly studied in the text. One such problem is that innocent agents might be given enough incentives by guilty agents to confess. For example, someone who is guilty might pay someone else who is innocent to take his place, or even worse, he might coerce him to. A related problem is the possibility of agents confessing to lesser crimes, rather than the ones they have committed. In this case, an innocent agent would still be confessing a crime he did not commit, but the difference is that he is guilty of committing a similar crime. 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 for criminal law would then have to find a way to resolve these type of problems in a satisfying manner. A way to, at least, mitigate these type of problems would be to "validate" the confession of a 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 Proof of Proposition 1

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

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

9.2 Proof of Proposition 4

Recall that the simplified nth agent problem is one of selecting $x_n(i, t_{-n}, \theta) \in [0, \phi]$ for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$, in order to maximize

$$\sum_{t_{-n}\in T_{-n}} \int_{\theta\in\Theta} \left(\pi\left(g, t_{-n}, \theta\right) - \alpha\pi\left(i, t_{-n}, \theta\right)\right) x_n\left(i, t_{-n}, \theta\right) d\theta$$

subject to

$$\sum_{t_{-n}\in T_{-n}} \int_{\theta\in\Theta} \frac{\pi\left(g, t_{-n}, \theta\right)}{\pi\left(t_n = g\right)} x_n\left(i, t_{-n}, \theta\right) d\theta \leq 1$$

Denote by $\widehat{\lambda}_n \geq 0$ the lagrange multiplier associated with the constraint and let $\widehat{\zeta}(t_{-n}, \theta) \geq 0$ and $\widehat{\eta}(t_{-n}, \theta) \geq 0$ be the multipliers associated with $x_n(i, t_{-n}, \theta) \leq \phi$ and $x_n(i, t_{-n}, \theta) \geq 0$ respectively. It follows that the optimal solution x_n^{SB} must be such that, for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$,

$$\pi\left(g, t_{-n}, \theta\right) - \alpha\pi\left(i, t_{-n}, \theta\right) + \widehat{\eta}\left(t_{-n}, \theta\right) = \widehat{\lambda}_n \frac{\pi\left(g, t_{-n}, \theta\right)}{\pi\left(t_n = g\right)} + \widehat{\zeta}\left(t_{-n}, \theta\right)$$

which can be written as

$$\pi\left(g, t_{-n}\right) l\left(\theta_n\right)\left(1 - \lambda_n\right) - \alpha\pi\left(i, t_{-n}\right) = \zeta\left(t_{-n}, \theta\right) - \eta\left(t_{-n}, \theta\right) \tag{9}$$

where $\lambda_n = \frac{\widehat{\lambda}_n}{\pi(t_n = g)}$, $\zeta(t_{-n}, \theta) = \frac{\widehat{\zeta}(t_{-n}, \theta)}{\pi(\theta_n | t_n = i) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}} | t_{\widetilde{n}})}$ and $\eta(t_{-n}, \theta) = \frac{\widehat{\eta}(t_{-n}, \theta)}{\pi(\theta_n | t_n = i) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}} | t_{\widetilde{n}})}$.

Notice that, for a fixed $t_{-n} \in T_{-n}$, the LHS is strictly increasing with θ_n, which means that there is a threshold $\theta_n^{SB}(t_{-n})$ such that

$$x_n(i, t_{-n}, \theta) = \begin{cases} \phi \text{ if } \theta_n \geq \theta_n^{SB}(t_{-n}) \\ 0 \text{ otherwise} \end{cases}$$

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

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

and so

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

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

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

which represents the expected punishment of the guilty agent under threshold $\theta_n^{SB}(t_{-n})$, given that he is indifferent between reporting truthfully and misreporting. Then, it follows that

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

where λ_n^* is such that $B_n(\phi, \lambda_n^*) = 1$. 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 $B_n\left(\overline{\phi}_n, 0\right) = 1$.

9.3 Proof of Proposition 5

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

$$B_n^g\left(x_n^{SB}\right) = \phi \sum_{t_{-n} \in T_{-n}} \frac{\pi(g, t_{-n})}{\pi(t_n = g)} \int_{\theta_n^{SB}(t_{-n})}^{1} \pi(\theta_n | t_n = g)\, d\theta_n = 1 \tag{10}$$

and

$$B_n^i\left(x_n^{SB}\right) = \phi \sum_{t_{-n} \in T_{-n}} \frac{\pi(i, t_{-n})}{\pi(t_n = i)} \int_{\theta_n^{SB}(t_{-n})}^{1} \pi(\theta_n | t_n = i)\, d\theta_n \tag{11}$$

Given (10) we have that (11) is equivalent to

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

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

9.4 Proof of Lemma 8

The goal of the proof is to show that, if there are two distinct messages with the same guiltiness ratio, it is possible to eliminate one of them. Therefore, it follows that it is possible to construct a mechanism where the messages sent all have distinct guiltiness ratios.

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

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

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

i) $\overline{\sigma}_n\left(t_n,m_n'\right)=\sigma_n\left(t_n,m_n'\right)+\sigma_n\left(t_n,m_n''\right)$ for $t_n=i,g$,

ii) $\overline{x}\left(m_n',m_{-n},\theta\right)=\frac{\sigma_n\left(t_n,m_n'\right)}{\sigma_n\left(t_n,m_n'\right)+\sigma_n\left(t_n,m_n''\right)}x\left(m_n',m_{-n},\theta\right)+\frac{\sigma_n\left(t_n,m_n''\right)}{\sigma_n\left(t_n,m_n'\right)+\sigma_n\left(t_n,m_n''\right)}x\left(m_n'',m_{-n},\theta\right)$ for $t_n=i,g$ and

iii) $\overline{x}\left(m_n'',m_{-n},\theta\right)=(1,...,1)$.

The new system merges the two messages and effectively eliminates message m''_n by making it undesirable to agent n. Notice that

$$B_n^{t_n}\left(\overline{x},\overline{\sigma}\right)=\frac{\sigma_n\left(t_n,m'_n\right)}{\sigma_n\left(t_n,m'_n\right)+\sigma_n\left(t_n,m''_n\right)}B_n^{t_n}\left(x,\sigma\right)+\frac{\sigma_n\left(t_n,m''_n\right)}{\sigma_n\left(t_n,m'_n\right)+\sigma_n\left(t_n,m''_n\right)}B_n^{t_n}\left(x,\sigma\right)=B_n^{t_n}\left(x,\sigma\right)$$

for $t_n=i,g$.

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

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

for some $m_{\widehat{n}}$ such that $\sigma_{\widehat{n}}\left(t_{\widehat{n}},m_{\widehat{n}}\right)>0$. Notice also that

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

Given that

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

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 $\overline{V}\left(\overline{x},\overline{\sigma}\right)=\overline{V}\left(x,\sigma\right)$.

The system $(\overline{x},\overline{\sigma})$ 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 remained unchanged. It is also renegotiation proof because, for all m_{-n},θ and for all $\widehat{n}$ (including n)

$$\overline{x}_{\widehat{n}}\left(m'_n,m_{-n},\theta\right)\leq\max\left\{x_{\widehat{n}}\left(m'_n,m_{-n},\theta\right),x_{\widehat{n}}\left(m''_n,m_{-n},\theta\right)\right\}\leq\gamma_{\widehat{n}}^{\sigma}\left(m'_n,m_{-n},\theta\right)=\gamma_{\widehat{n}}^{\overline{\sigma}}\left(m'_n,m_{-n},\theta\right)$$

Finally, because for all m_n and for all n, $\frac{\sigma_n^{RP}(g,m_n)}{\sigma_n^{RP}(i,m_n)}\in\mathbb{R}_+\cup\{\infty\}$, the statement follows where I simply denote messages that are sent only by guilty agents as $\{c\}$.

9.5 Proof of Lemma 9

First, I start by showing 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 than that of the guilty agent.

Notice that it is possible to write

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

Define $E_n \equiv \left\{ (m_{-n}, \theta) \in M_{-n} \times [0,1]^N : x_n^\sigma (m_n^\sigma, m_{-n}, \theta) = 1 \right\}$. 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 $\int\limits_{e_n \in E_n} \pi(e_n|t_n = g)\, de_n > \int\limits_{e_n \in E_n} \pi(e_n|t_n = i)\, de_n$ and so the expected punishment of the agent when sending message m_n^σ is higher if he is guilty. Finally, if there is $e_n' \in E_n$ such that $\frac{\pi\left(e_n'|t_n=i\right)}{\pi\left(e_n'|t_n=g\right)} \geq 1$ and given that $x_n^\sigma (m_n^\sigma, m_{-n}, \theta)$ 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, $\int\limits_{e_n \notin E_n} \pi(e_n|t_n = g)\, de_n < \int\limits_{e_n \notin E_n} \pi(e_n|t_n = i)\, de_n$ which implies that $\int\limits_{e_n \in E_n} \pi(e_n|t_n = g)\, de_n > \int\limits_{e_n \in E_n} \pi(e_n|t_n = i)\, de_n$ and so, also in this case, the expected punishment of the agent when sending message m_n^σ is higher if he is guilty. Hence, it follows that the system (x^σ, σ) is incentive compatible.

To guarantee the system is renegotiation proof I set the beliefs after any message that is not sent in equilibrium to be as if the agent's "guiltiness" ratio is equal to 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 $\gamma_n^\sigma (m_n^\sigma, m_{-n}, \theta) \leq \gamma_n^\sigma (m_n, m_{-n}, \theta)$ for all $m_n \in \mathbb{R}_+$.

Now, I show that there is no other incentive compatible and renegotiation proof system that induces a strictly higher expected utility for the principal, i.e. for all σ and for any $x : M \times \Theta \to \mathbb{R}_+^N$, $\widehat{V}(x^\sigma, \sigma) \geq \widehat{V}(x, \sigma)$.

Take any system (x, σ) and assume it is incentive compatible and renegotiation proof. Now consider the alternative system (x', σ) such that, for all m_{-n}, θ and n,
i) $x_n' (c, m_{-n}, \theta) = x_n (c, m_{-n}, \theta)$ and
ii) $x_n' (m_n, m_{-n}, \theta) = \gamma_n^\sigma (m_n^\sigma, m_{-n}, \theta)$ for all $m_n \in \mathbb{R}_+$.

Notice that one can write

$$\widehat{V}_n (x', \sigma) = \int\limits_{m \in M} \int\limits_{\theta \in \Theta} \pi^\sigma (m, \theta)\, \kappa_n^\sigma (m, \theta, x_n' (m, \theta))\, d\theta dm$$

where $\pi^\sigma (m, \theta)$ denotes the joint density of m and θ, given strategy profile σ and

$$\kappa_n^\sigma (m, \theta, x_n' (m, \theta)) = (\pi^\sigma (t_n = g|m, \theta) - \pi^\sigma (t_n = i|m, \theta))\, x_n' (m, \theta)$$

Given that, by definition,

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

and because $\kappa_n^\sigma (m, \theta, \cdot)$ is single peaked around $\gamma_n^\sigma (m, \theta)$, we have that

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

Hence, it follows that $\widehat{V}(x', \sigma) \geq \widehat{V}(x, \sigma)$. However, (x', σ) may not be incentive compatible given that the punishments after non-confessing messages have increased with respect (x, σ) but punishments after a confession stayed the same.

Finally, compare (x^σ, σ) with (x', σ). Notice that the expected punishment after sending non-confessing messages is equal in both system, so to the difference between the two lies on the fact that punishments after confessions are higher in x^{σ} in order to satisfy incentive compatibility. Hence, it must be that

$$
\begin{aligned}
& \int_{m_{-n}} \int_{\theta} \pi^{\sigma}\left(c, m_{-n}, \theta\right) \kappa_n^{\sigma}\left(c, m_{-n}, \theta, x_n^{\sigma}\left(c, m_{-n}, \theta\right)\right) d\theta dm_{-n} \\
\geq & \int_{m_{-n}} \int_{\theta} \pi^{\sigma}\left(c, m_{-n}, \theta\right) \kappa_n^{\sigma}\left(c, m_{-n}, \theta, x_n^{\prime}\left(c, m_{-n}, \theta\right)\right) d\theta dm_{-n}
\end{aligned}
$$

and so $\widehat{V}\left(x^{\sigma}, \sigma\right) \geq \widehat{V}\left(x^{\prime}, \sigma\right) \geq \widehat{V}\left(x, \sigma\right)$.

9.6 Proof of Proposition 10

Let σ^{CIS} denote the optimal CIS, i.e. $\widehat{V}\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right) \geq \widehat{V}\left(x^{\sigma}, \sigma\right)$ for all $\sigma \in \Phi^{CIS}$, where

$$
\Phi^{CIS} = \left\{\sigma \in \Phi : \sigma_n\left(t_n, m_n\right) = 0 \text{ for all } m_n \neq \left\{c, m_n^{\sigma}\right\} \text{ and for all } t_n \text{ and } n \ \right\}
$$

Suppose the statement is not true. Then there is another system $\left(x^{\widetilde{\sigma}}, \widetilde{\sigma}\right)$ that is not a CIS that is strictly preferred to $\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right)$. This means that under $\widetilde{\sigma}$ there is at least one agent that sends a second non-confessing message. Without loss of generality, assume that agent 1 is the agent that sends this second non-confessing message $m_1^{\prime} \notin \left\{c, m_1^{\widetilde{\sigma}}\right\}$. In particular, assume that $r_1\left(m_1^{\widetilde{\sigma}}\right) < r_1\left(m_1^{\prime}\right) < \infty$ because, otherwise, by the logic of Lemma 8, the proposition would follow trivially.

Consider system $\left(x^{\widehat{\sigma}}, \widehat{\sigma}\right)$ where $\widehat{\sigma} = \widetilde{\sigma}$ except that $\widehat{\sigma}_1\left(g, m_1^{\prime}\right) = \widetilde{\sigma}_1\left(g, m_1^{\prime}\right) - v$, $\widehat{\sigma}_1\left(g, c\right) = \widetilde{\sigma}_1\left(g, c\right) + v$, where v is such that

$$
\frac{\widetilde{\sigma}_1\left(i, m_1^{\prime}\right)}{\widetilde{\sigma}_1\left(g, m_1^{\prime}\right) - v} = \frac{\widetilde{\sigma}_1\left(i, m_1^{\widetilde{\sigma}}\right)}{\widetilde{\sigma}_1\left(g, m_1^{\widetilde{\sigma}}\right)}
$$

I show that system $\left(x^{\widehat{\sigma}}, \widehat{\sigma}\right)$ is strictly preferred to system $\left(x^{\widetilde{\sigma}}, \widetilde{\sigma}\right)$ that is a contradiction with $\left(x^{\widetilde{\sigma}}, \widetilde{\sigma}\right)$ being optimal and so shows the statement of the proposition.

Write $\overline{V}\left(\sigma\right) = \widehat{V}\left(x^{\sigma}, \sigma\right)$ for all σ. Notice that $\overline{V}_1\left(\widehat{\sigma}\right) = \overline{V}_1\left(\widetilde{\sigma}\right)$. It also follows that, for all n,

$$
\begin{aligned}
& \overline{V}_n\left(\widehat{\sigma}\right) - \overline{V}_n\left(\widetilde{\sigma}\right) \\
= & \int_{m_n, m_{-1,n}, \theta} \sum_{t_{-1,n} \in T_{-1,n}} B \left(\prod_{\widetilde{n} \neq 1, n} \widetilde{\sigma}_{\widetilde{n}}\left(t_{\widetilde{n}}, m_{\widetilde{n}}\right) \pi\left(\theta_{\widetilde{n}} | t_{\widetilde{n}}\right) \right) d\left(m_n, m_{-1,n}, \theta\right)
\end{aligned}
$$

where

$$
\begin{aligned}
B = & \ v \begin{pmatrix} \pi\left(g, g, t_{-1,n}\right) \widetilde{\sigma}_n\left(g, m_n\right) \pi\left(\theta_n | t_1 = g\right) \pi\left(\theta_1 | t_1 = g\right) \\ -\alpha \pi\left(i, g, t_{-1,n}\right) \widetilde{\sigma}_n\left(i, m_n\right) \pi\left(\theta_n | t_n = i\right) \pi\left(\theta_1 | t_1 = g\right) \end{pmatrix} \gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, c, m_{-1,n}\right), \theta\right) + \\
& \begin{pmatrix} \pi\left(g, g, t_{-1,n}\right) \widetilde{\sigma}_n\left(g, m_n\right) \pi\left(\theta_n | t_n = g\right) \left(\widetilde{\sigma}_1\left(g, m_1^{\prime}\right) - v\right) \pi\left(\theta_1 | t_1 = g\right) \\ +\pi\left(g, i, t_{-1,n}\right) \widetilde{\sigma}_n\left(g, m_n\right) \pi\left(\theta_n | t_n = g\right) \widetilde{\sigma}_1\left(i, m_1^{\prime}\right) \pi\left(\theta_1 | t_1 = i\right) - \\ \alpha \pi\left(i, g, t_{-1,n}\right) \widetilde{\sigma}_n\left(i, m_n\right) \pi\left(\theta_n | t_n = i\right) \left(\widetilde{\sigma}_1\left(g, m_1^{\prime}\right) - v\right) \pi\left(\theta_1 | t_1 = g\right) \\ -\alpha \pi\left(i, i, t_{-1,n}\right) \widetilde{\sigma}_n\left(i, m_n\right) \pi\left(\theta_n | t_n = i\right) \widetilde{\sigma}_1\left(i, m_1^{\prime}\right) \pi\left(\theta_1 | t_1 = i\right) \end{pmatrix} \gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, m_1^{\widetilde{\sigma}}, m_{-1,n}\right), \theta\right) \\
& - \begin{pmatrix} \pi\left(g, g, t_{-1,n}\right) \widetilde{\sigma}_n\left(g, m_n\right) \pi\left(\theta_n | t_n = g\right) \widetilde{\sigma}_1\left(g, m_1^{\prime}\right) \pi\left(\theta_1 | t_1 = g\right) \\ +\pi\left(g, i, t_{-1,n}\right) \widetilde{\sigma}_n\left(g, m_n\right) \pi\left(\theta_n | t_n = g\right) \widetilde{\sigma}_1\left(i, m_1^{\prime}\right) \pi\left(\theta_1 | t_1 = i\right) \\ -\alpha \pi\left(i, g, t_{-1,n}\right) \widetilde{\sigma}_n\left(i, m_n\right) \pi\left(\theta_n | t_n = i\right) \widetilde{\sigma}_1\left(g, m_1^{\prime}\right) \pi\left(\theta_1 | t_1 = g\right) \\ -\alpha \pi\left(i, i, t_{-1,n}\right) \widetilde{\sigma}_n\left(i, m_n\right) \pi\left(\theta_n | t_n = i\right) \widetilde{\sigma}_1\left(i, m_1^{\prime}\right) \pi\left(\theta_1 | t_1 = i\right) \end{pmatrix} \gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, m_1^{\prime}, m_{-1,n}\right), \theta\right)
\end{aligned}
$$

Notice that by replacing $\gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, m_1^{\widetilde{\sigma}}, m_{-1,n}\right),\theta\right)$ by $\gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, m_1', m_{-1,n}\right),\theta\right)$ in the second line, it is possible to write that

$$B > v\begin{pmatrix} \pi\left(g,g,t_{-1,n}\right)\widetilde{\sigma}_n\left(g,m_n\right)\pi\left(\theta_n|t_1=g\right)\pi\left(\theta_1|t_1=g\right) \\ -\alpha\pi\left(i,g,t_{-1,n}\right)\widetilde{\sigma}_n\left(i,m_n\right)\pi\left(\theta_n|t_n=i\right)\pi\left(\theta_1|t_1=g\right)\end{pmatrix} * \begin{pmatrix} \gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, c, m_{-1,n}\right),\theta\right) \\ -\gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}}, m_1', m_{-1,n}\right),\theta\right)\end{pmatrix} > 0$$

by the definition of $m_n^{\widetilde{\sigma}}$ and $\gamma_n^{\widetilde{\sigma}}(m,\theta)$ for all (m,θ) and n. This implies that, for all n, $\overline{V}_n(\widehat{\sigma}) > \overline{V}_n(\widetilde{\sigma})$ which completes the proof.[13]

9.7 Proof of Proposition 11

In a CIS, only two messages are sent: c and m_n^{σ} for each agent n. Denote the optimal CIS by $\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right)$ and let $\tau \in [0,1]^N$ be such that $\sigma_n^{CIS}\left(g, m_n^{\sigma}\right) = \tau_n$ for all n. Also, let $\overline{V}(\tau)$ denote the corresponding expected utility of the principal. A trial system is characterized by $\tau = \underline{\tau} \equiv (1, ..., 1)$.

I show the statement by showing that a) $\frac{\partial \overline{V}_n}{\partial \tau_n}(\underline{\tau}) = 0$ for all n and b) $\lim_{\tau_n \to 1} \frac{\partial \overline{V}_{\widehat{n}}}{\partial \tau_n}(1, ..., \tau_n, ...1) \leq 0$ for all n and $\widehat{n}$, with the inequality being strict for at least one pair $(\widehat{n}, n)$, unless the types of the agents are independent.

Notice that it is possible to write $\overline{V}(\tau) = \sum\limits_{n=1}^{N} \overline{V}_n(\tau)$ where

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

where

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

and

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

The threshold $\theta_n^{CIS}\left(m_{-n}, \theta_{-n}\right)$ is such that $x_n^{CIS}\left(m_n^{\sigma^{CIS}}, m_{-n}, \theta_{-n}\right) = \begin{cases} 0 \text{ if } \theta_n \leq \theta_n^{CIS}\left(m_{-n}, \theta_{-n}\right) \\ \quad 1 \text{ otherwise}\end{cases}$.

As for a), notice that

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

Given that, when $\tau_n = 1$,

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

[13] Recall that $\gamma_n^{\sigma}(m,\theta) \in \arg\max_{x\in[0,1]} \left\{\left(\pi\left(t_n = g|m,\theta\right) - \alpha\pi\left(t_n = g|m,\theta\right)\right)x\right\}$ which is equal to $\arg\max_{x\in[0,1]} \left\{\left(\pi\left(t_n = g, m,\theta\right) - \alpha\pi\left(t_n = g, m,\theta\right)\right)x\right\}$.

it must be that $\frac{\partial \overline{V}_n}{\partial \tau_n}(\underline{\tau}) = 0$.

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

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

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

$$\int\limits_{m_{-\widehat{n},n} \in M_{-\widehat{n},n}} \int\limits_{\theta_{-\widehat{n}} \in \Theta_{-\widehat{n}}} \left[\begin{array}{c} -A^{CIS}\left(\theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right), \theta_{-\widehat{n}}, m_n = c, m_{-\widehat{n},n}\right) * \\ \frac{d\theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}{d\tau_n} \\ -A^{CIS}\left(\theta_{\widehat{n}}^{CIS}\left(m_n = m_n^{\sigma^{CIS}}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right), \theta_{-\widehat{n}}, m_n = m_n^{\sigma^{CIS}}, m_{-\widehat{n},n}\right) * \\ \frac{d\theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}{d\tau_n} \end{array} \right] d\theta_{-\widehat{n}} dm_{-\widehat{n},n} \tag{12}$$

$$+ \int\limits_{m_{-\widehat{n},n} \in M_{-\widehat{n},n}} \int\limits_{\theta_{-\widehat{n}} \in \Theta_{-\widehat{n}}} \left[\begin{array}{c} \int\limits_{\theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}^{1} \frac{dA^{CIS}\left(\theta_{\widehat{n}}, \theta_{-\widehat{n}}, m_n = c, m_{-\widehat{n},n}\right)}{d\tau_n} d\theta_{\widehat{n}} \\ + \int\limits_{\theta_{\widehat{n}}^{CIS}\left(m_n = m_n^{\sigma^{CIS}}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}^{1} \frac{dA^{CIS}\left(\theta_{\widehat{n}}, \theta_{-\widehat{n}}, m_n = m_n^{\sigma^{CIS}}, m_{-\widehat{n},n}\right)}{d\tau_n} d\theta_{\widehat{n}} \end{array} \right] d\theta_{-\widehat{n}} dm_{-\widehat{n},n} \tag{13}$$

Notice that (12) is equal to 0 when $\tau_{\widehat{n}} = 1$ given that

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

Let

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

and

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

Then, condition (13) can be written as

$$\begin{aligned} &\int\limits_{m_{-\widehat{n},n} \in M_{-\widehat{n},n}} \int\limits_{\theta_{-\widehat{n}} \in \overline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}}} \int\limits_{\theta_{\widehat{n}}^{SCI}\left(m_n = m_n^{\sigma^{CIS}}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}^{\theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)} B^{CIS}\left(\theta_{\widehat{n}}, \theta_{-\widehat{n}}, m_{-\widehat{n},n}\right) d\theta_{\widehat{n}} d\theta_{-\widehat{n}} dm_{-\widehat{n},n} \\ &- \int\limits_{m_{-\widehat{n},n} \in M_{-\widehat{n},n}} \int\limits_{\theta_{-\widehat{n}} \in \underline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}}} \int\limits_{\theta_{\widehat{n}}^{SCI}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}^{\theta_{\widehat{n}}^{CIS}\left(m_n = m_n^{\sigma^{CIS}}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)} B^{CIS}\left(\theta_{\widehat{n}}, \theta_{-\widehat{n}}, m_{-\widehat{n},n}\right) d\theta_{\widehat{n}} d\theta_{-\widehat{n}} dm_{-\widehat{n},n} \end{aligned}$$

where

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

that is strictly negative when $\tau_{\widehat{n}} = 1$, 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.

9.8 Proof of Proposition 14

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

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

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

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

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

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

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

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

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

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

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 $\theta_{-n} \in \Theta_{-n}$, $\lim_{\theta_n \to 0} x_n^{Tr}\left(\left(\theta_n, \theta_{-n}\right)\right) = 0$ and given that $\lim_{\theta_n \to 1} l(\theta_n) = \infty$ it must be that, for all $\theta_{-n} \in \Theta_{-n}$, $\lim_{\theta_n \to 1} x_n^{Tr}\left(\left(\theta_n, \theta_{-n}\right)\right) = 1$.

9.9 Proposition 15

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

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

subject

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

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

$$\pi\left(g, t_{-n}, \theta\right) \frac{\partial u_{n}^{p}\left(g, x_{n}\left(g, t_{-n}, \theta\right)\right)}{\partial x_{n}} + \lambda_{n} \pi\left(g, t_{-n}, \theta\right) \frac{\partial u^{g}\left(g, x_{n}\left(g, t_{-n}, \theta\right)\right)}{\partial x_{n}} = \zeta_{n}^{g}\left(t_{-n}, \theta\right) - \eta_{n}^{g}\left(t_{-n}, \theta\right)$$

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

Given that

$$\frac{\partial^{2} u_{n}^{p}\left(g, \cdot\right)}{\partial\left(x_{n}\right)^{2}} + \lambda_{n} \frac{\partial^{2} u^{g}\left(g, \cdot\right)}{\partial\left(x_{n}\right)^{2}} < 0$$

and

$$\frac{\partial u_{n}^{p}\left(g, 1\right)}{\partial x_{n}} + \lambda_{n} \frac{\partial u^{g}\left(g, 1\right)}{\partial x_{n}} < 0$$

and

$$\frac{\partial u_{n}^{p}\left(g, 0\right)}{\partial x_{n}} + \lambda_{n} \frac{\partial u^{g}\left(g, 0\right)}{\partial x_{n}} > 0$$

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

$$\frac{\partial u_{n}^{p}\left(g, \widetilde{x}_{n}^{SB}\left(g, t_{-n}, \theta\right)\right)}{\partial x_{n}} + \lambda_{n} \frac{\partial u^{g}\left(g, \widetilde{x}_{n}^{SB}\left(g, t_{-n}, \theta\right)\right)}{\partial x_{n}} = 0$$

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

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

because the incentive constraint binds.

9.10 Proof of Proposition 16

The first order condition with respect to $x_n\left(i, t_{-n}, \theta\right)$ is given by

$$\alpha \pi\left(i, t_{-n}, \theta\right) \frac{\partial u^{i}\left(x_{n}\left(i, t_{-n}, \theta\right)\right)}{\partial x_{n}} - \lambda_{n} \pi\left(g, t_{-n}, \theta\right) \frac{\partial u^{g}\left(x_{n}\left(i, t_{-n}, \theta\right)\right)}{\partial x_{n}} = \zeta_{n}^{g}\left(t_{-n}, \theta\right) - \eta_{n}^{g}\left(t_{-n}, \theta\right)$$

which can be written as

$$\begin{aligned} & -\alpha \pi\left(i, t_{-n}\right) \pi\left(\theta_{n} | t_{n} = i\right) \omega_{i}\left(x_{n}\left(i, t_{-n}, \theta\right)\right)^{\omega_{i}-1} + \lambda_{n} \pi\left(g, t_{-n}\right) \pi\left(\theta_{n} | t_{n} = g\right) \omega_{g}\left(x_{n}\left(i, t_{-n}, \theta\right)\right)^{\omega_{g}-1} \\ = & \frac{\zeta_{n}^{g}\left(t_{-n}, \theta\right) - \eta_{n}^{g}\left(t_{-n}, \theta\right)}{\pi\left(\theta_{-n} | t_{-n}\right)} \end{aligned}$$

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

$$\psi_{n}\left(t_{-n}, \theta_{n}\right) = \left(\frac{\lambda_{n} \omega_{g}}{\alpha \omega_{i}} \frac{\pi\left(g, t_{-n}\right)}{\pi\left(i, t_{-n}\right)} l\left(\theta_{n}\right)\right)^{\frac{1}{\omega_{i} - \omega_{g}}}$$

Notice that

$$\alpha\pi\left(i,t_{-n},\theta\right)\frac{\partial^{2}u^{i}\left(\psi_{n}\left(t_{-n},\theta_{n}\right)\right)}{\partial\left(x_{n}\right)^{2}}-\lambda_{n}\pi\left(g,t_{-n},\theta\right)\frac{\partial^{2}u^{g}\left(\psi_{n}\left(t_{-n},\theta_{n}\right)\right)}{\partial\left(x_{n}\right)^{2}}$$

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

9.11 Proof of Proposition 17

An optimal allocation 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 an allocation $x:\widehat{T}\times\Theta\rightarrow[0,1]^N$ 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

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

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

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

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

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

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.

9.12 Proof of Proposition 18

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

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

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

or B) for all $\left(t_n, \beta_n\right)$,

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

Proof of Lemma 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 $s = \left\{ \{s_n(t_n, \beta_n)\}_{\beta_n \in [0,1]} \right\}_{t_n \in T_n}$ and $x : \{T_n \times [0,1]\}_{n=1}^{N} \times \Theta \to [0,1]$. For all n, let $B_n^{t_n}(\beta_n)$ 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 $\overline{c}$, $\Gamma_{\overline{c}}^{t_n}$ 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

$$\begin{aligned}
&\pi(t_n = g)\,\pi(\beta_n \in \Gamma_c^g \cup \Gamma_{=}^g | t_n = g)\, x_n(c) - \alpha\pi(t_n = i)\,\pi\left(\beta_n \in \Gamma_c^i \cup \Gamma_{=}^i | t_n = i\right) x_n(c) \\
&+ \int\limits_{\beta_n \in \Gamma_{\overline{c}}^g} \int\limits_{\beta_{-n}} \int\limits_{\theta} \sum_{t_{-n}} \sum_{m_{-n}} \pi(g, t_{-n})\,\pi(\beta | t_n = g, t_{-n})\,\pi(m_{-n}, \theta | t_n = g, t_{-n}, \beta)\, x_n(\overline{c}, m_{-n}, \theta)\, d\theta d\beta \\
&-\alpha \int\limits_{\beta_n \in \Gamma_{\overline{c}}^i} \int\limits_{\beta_{-n}} \int\limits_{\theta} \sum_{t_{-n}} \sum_{m_{-n}} \pi(i, t_{-n})\,\pi(\beta | t_n = i, t_{-n})\,\pi(m_{-n}, \theta | t_n = i, t_{-n}, \beta)\, x_n(\overline{c}, m_{-n}, \theta)\, d\theta d\beta
\end{aligned}$$

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

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

and for all $\beta_n \in \Gamma_{=}^i$,

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

For all $\beta_n \in \Gamma_{=}^g$ and $\beta_n \in \Gamma_{=}^i$ let $\lambda^g(\beta_n) \geq 0$ and $\lambda^i(\beta_n) \geq 0$ denote the lagrange multipliers of the conditions above respectively. Also, for all $\beta_n \in \Gamma_{\overline{c}}^g$ and $\beta_n \in \Gamma_{\overline{c}}^i$, write $\lambda^g(\beta_n) = \lambda^i(\beta_n) = 1$.

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

$$\begin{aligned}
&\int\limits_{\beta_n \in \Gamma_{=}^g \cup \Gamma_{\overline{c}}^g} \pi(\beta_n | t_n = g)\,\lambda^g(\beta_n) \int\limits_{\beta_{-n}} \sum_{t_{-n}} \pi(g, t_{-n})\,\pi\left(\beta_{-n} | t_{-n}\right)\pi\left(m_{-n}, \theta | t_{-n}, \beta_{-n}\right) d\beta \\
&-\alpha \int\limits_{\beta_n \in \Gamma_{=}^i \cup \Gamma_{\overline{c}}^i} \pi(\beta_n | t_n = i)\,\lambda^i(\beta_n) \int\limits_{\beta_{-n}} \sum_{t_{-n}} \pi(i, t_{-n})\,\pi\left(\beta_{-n} | t_{-n}\right)\pi\left(m_{-n}, \theta | t_{-n}, \beta_{-n}\right) d\beta \\
= \quad &\zeta_n^{\overline{c}}(m_{-n}, \theta) - \eta_n^{\overline{c}}(m_{-n}, \theta)
\end{aligned}$$

where $\zeta_n^{\overline{c}}(m_{-n}, \theta) \geq 0$ and $\eta_n^{\overline{c}}(m_{-n}, \theta) \geq 0$ denote the lagrange multipliers associated with constraints $\{x_n(\overline{c}, m_{-n}, \theta) \geq 0\}$ and $\{x_n(\overline{c}, m_{-n}, \theta) \leq 1\}$ 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

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

and

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

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, for all m_{-n}, 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 B) cannot be.

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

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

This implies that

$$\frac{\int_{\beta_n^i}^{1} \pi(\beta_n | t_n = i)\, \beta_n d\beta_n}{\int_{\beta_n^i}^{1} \pi(\beta_n | t_n = i)\, d\beta_n} > \frac{\int_{\beta_n^g}^{1} \pi(\beta_n | t_n = g)\, \beta_n d\beta_n}{\int_{\beta_n^g}^{1} \pi(\beta_n | t_n = g)\, d\beta_n}$$

where, without loss of generality, $\lambda^{t_n}(\beta_n) = 1$ for all t_n and β_n because there is two pairs (i, β_n^i) and (g, β_n^g) 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 go to trial (or is indifferent). I do this by showing that, for any fixed β_n, the expected punishment of going to trial is higher if the agent is guilty. The proof is the analogous to the one of Lemma 9. Notice that

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

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

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

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

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

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 $\int\limits_{e_n \notin E_n^{\theta_n}} \pi(e_n | t_n = i)\, de_n > \int\limits_{e_n \notin E_n^{\theta_n}} \pi(e_n | t_n = g)\, de_n$ which implies the statement.

9.13 Proof of Proposition 19

Suppose the principal waits until he receives evidence θ and then makes a proposal $y_\theta : T \times \Theta \to R_+^N$ such that it is a Bayes-Nash equilibrium for all agents to tell the truth. We will show that $x_y(t, \theta)$ satisfies (IC) - the relevant incentive constraint when the principal acts before observing the evidence.

Given each proposal y_θ and their type own t_n, agents form some posterior belief about t and θ whose joint density we denote by $\pi^{y_\theta}(t,\theta|t_n)$. Given that y_θ is incentive compatible 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,

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

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[14]. 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

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

By following the same steps with the right hand side, condition (IC) follows.

9.14 Proof of Proposition 20

I implement allocation x^{IP} by considering strategy y for the principal where $y_{\widehat{\theta}}(t,\theta) = x_y(t,\theta)$ for all $\widehat{\theta} \in \Theta$. I start by specifying beliefs in case the principal proposes a different mechanism than x^{IP}. Given that such a proposal is off the equilibrium path, I have the freedom to specify any beliefs for the agents. Hence, I set the agents' beliefs to be such that, whenever any other proposal is made, the agents believe that $(0,...,0)$ is the realized θ with probability 1. This means that each agent will put probability 1 in every other agent being guilty, regardless of their own type, which implies that, for the deviation proposal $\widehat{y}_{\widehat{\theta}}$ to be incentive compatible for some $\widehat{\theta}$, it must be that, for all n, $\widehat{y}_{\widehat{\theta},n}(i,(g,..,g),(0,..,0)) = \widehat{y}_{\theta,n}(g,(g,..,g),(0,..,0))$. As for $\widehat{y}_{\widehat{\theta},n}(t,\theta)$ for all other t and θ it is irrelevant as the agents will put no weight into these events occurring.

It follows the maximum deviation payoff the principal can get from each agent n, given the observed $\widehat{\theta}$, is

$$\max_{\beta\in[0,1]}\left\{\left(\sum_{t_{-n}\in T_{-n}}\pi\left(g,t_{-n}|\widehat{\theta}\right) - \alpha\sum_{t_{-n}\in T_{-n}}\pi\left(i,t_{-n}|\widehat{\theta}\right)\right)\beta\right\} - \sum_{t_{-n}\in T_{-n}}\pi\left(g,t_{-n}|\widehat{\theta}\right)$$

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

By definition of x^{IP}, it follows that the payoff of proposing x^{IP} for a given $\widehat{\theta}$ is given by

$$\sum_{t_{-n}\in T_{-n}} \max_{\beta\in[0,1]} \left\{ \left(\pi\left(g, t_{-n}|\widehat{\theta}\right) - \alpha\pi\left(i, t_{-n}|\widehat{\theta}\right)\right)\beta \right\} - \sum_{t_{-n}\in T_{-n}} \pi\left(g, t_{-n}|\widehat{\theta}\right)$$

Given that

$$\begin{aligned} & \sum_{t_{-n}\in T_{-n}} \max_{\beta\in[0,1]} \left\{ \left(\pi\left(g, t_{-n}|\widehat{\theta}\right) - \alpha\pi\left(i, t_{-n}|\widehat{\theta}\right)\right)\beta \right\} \\ \geq\ & \max_{\beta\in[0,1]} \left\{ \left(\sum_{t_{-n}\in T_{-n}} \pi\left(g, t_{-n}|\widehat{\theta}\right) - \alpha \sum_{t_{-n}\in T_{-n}} \pi\left(i, t_{-n}|\widehat{\theta}\right)\right)\beta \right\} \end{aligned}$$

the principal has no incentive to deviate.

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

[5] 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.

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

[7] 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.

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

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

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

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

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

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

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

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

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

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

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

[19] Myerson, Roger B. "Incentive compatibility and the bargaining problem." Econometrica: journal of the Econometric Society (1979): 61-73.

[20] Myerson, Roger B. "Mechanism design by an informed principal." Econometrica: Journal of the Econometric Society (1983): 1767-1797.

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

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

[23] Santobello v. New York, 404 U.S. 257 (1971), 261

[24] Scott, Robert E., and William J. Stuntz. "Plea bargaining as contract." Yale Law Journal (1992): 1909-1968.

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

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

[27] White, Welsh S. "Police trickery in inducing confessions." University of Pennsylvania Law Review (1979): 581-629.