2016

Essays in Applied Microeconomic Theory

Francisco Silva
University of Pennsylvania, fsilva@sas.upenn.edu

Essays in Applied Microeconomic Theory

Abstract

This dissertation consists of three chapters that each study one applied microeconomic theory problem. In the first chapter, I consider the problem a social planner faces in constructing a criminal justice system which addresses two needs: to protect the innocent and to punish the guilty. I characterize the socially optimal criminal justice system under various assumptions with respect to the social planner's ability to commit. In the optimal system, before a criminal investigation is initiated, all members of the community are given the opportunity to confess to having committed the crime in exchange for a smaller than socially optimal punishment, which is independent of any future evidence that might be discovered. Agents who choose not to confess might be punished once the investigation is completed if the evidence gathered is sufficiently incriminatory. In this paper's framework, leniency for confessing agents is efficient not because it saves resources or reduces risk, but because there are informational 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. \par

In the second chapter, I present a theory which shows how the influence of others may generate overconfidence. The argument is built on the idea that the more help an agent receives when performing a task, the less informative the score on that task will be relative to the agent's ability to perform it. Overconfident agents, who tend to benefit from more cooperation opportunities simply because they are perceived to be more skilled, will remain overconfident because the future signals they will observe will contain very little information regarding their ability. On the contrary, the scores on tasks that underconfident agents receive will be more informative, which will help them learn their true ability faster. \par

Finally, in the third chapter, I compare two different systems of provision of discrete public goods: a centralized system, ruled by a benevolent dictator who has no commitment power; and an anarchic system, based on voluntary contributions, where there is no ruler. If the public good is binary, then the public good provision problem is merely an informational one. In this environment, I show that the anarchic system can always replicate any outcome of the centralized system. However, as the number of alternatives available increases, the classical free riding problem described in Samuelson (1954) emerges. As the classical free riding problem becomes more important relative to the informational free riding problem, the centralized system becomes the preferred system of the two.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Economics

First Advisor

Andrew Postlewaite

Subject Categories

Economics

ESSAYS IN APPLIED MICROECONOMIC THEORY

Francisco Silva

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Andrew Postlewaite, Professor

Graduate Group Chairperson

Jesus Fernandez-Villaverde, Professor

Dissertation Committee

Andrew Postlewaite, Professor of Economics

Steven A. Matthews, Professor of Economics

Rakesh Vohra, Professor of Economics

ACKNOWLEDGEMENTS

I would like to thank David Abrams, Diego Amador, Lorenzo Braccini, Mustafa Dogan, Selman Erol, Hanming Fang, Tzuo-Hann Law, Anqi Li, Sangmok Lee, Nicholas Janetos, George Mailath, Steven A. 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.

ABSTRACT

ESSAYS IN APPLIED MICROECONOMIC THEORY

Francisco Silva

Andrew Postlewaite

This dissertation consists of three chapters that each study one applied microeconomic theory problem. In the first chapter, I consider the problem a social planner faces in constructing a criminal justice system which addresses two needs: to protect the innocent and to punish the guilty. I characterize the socially optimal criminal justice system under various assumptions with respect to the social planner's ability to commit. In the optimal system, before a criminal investigation is initiated, all members of the community are given the opportunity to confess to having committed the crime in exchange for a smaller than socially optimal punishment, which is independent of any future evidence that might be discovered. Agents who choose not to confess might be punished once the investigation is completed if the evidence gathered is sufficiently incriminatory. In this paper's framework, leniency for confessing agents is efficient not because it saves resources or reduces risk, but because there are informational 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.

In the second chapter, I present a theory which shows how the influence of others may generate overconfidence. The argument is built on the idea that the more help an agent receives when performing a task, the less informative the score on that task will be relative to the agent's ability to perform it. Overconfident agents, who tend to benefit from more cooperation opportunities simply because they are perceived to be more skilled, will remain overconfident because the future signals they will observe will contain very little information regarding their ability. On the contrary, the scores on tasks that underconfident

agents receive will be more informative, which will help them learn their true ability faster.

Finally, in the third chapter, I compare two different systems of provision of discrete public goods: a centralized system, ruled by a benevolent dictator who has no commitment power; and an anarchic system, based on voluntary contributions, where there is no ruler. If the public good is binary, then the public good provision problem is merely an informational one. In this environment, I show that the anarchic system can always replicate any outcome of the centralized system. However, as the number of alternatives available increases, the classical free riding problem described in Samuelson (1954) emerges. As the classical free riding problem becomes more important relative to the informational free riding problem, the centralized system becomes the preferred system of the two.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

ABSTRACT iii

LIST OF TABLES vii

LIST OF ILLUSTRATIONS viii

CHAPTER 1 : If we Confess our Sins 1

1.1 Introduction 1

1.2 Related Literature 9

1.3 Model 12

1.4 Trial System 15

1.5 Second Best Problem 16

1.6 Limited Commitment Power 31

1.7 Extensions 41

1.8 Concluding remarks 56

CHAPTER 2 : Inducing Overconfidence 59

2.1 Introduction 59

2.2 The Model 62

2.3 Induced overconfidence 63

2.4 Why is help increasing? 67

2.5 Conclusion 72

CHAPTER 3 : Should the Government provide public goods if it cannot commit? 73

3.1 Introduction 73

3.2 Model 79

3.3 Benevolent Dictator 79
3.4 Anarchy 90
3.5 Discrete Public Good - general case 95
3.6 Extensions 97
3.7 Conclusion 101

APPENDIX 104
A.1 Appendix to Chapter 1 104
A.2 Appendix to Chapter 2 132
A.3 Appendix to Chapter 3 140

BIBLIOGRAPHY 153

LIST OF TABLES

TABLE 1 : Prior Distribution of Guilt 22

LIST OF ILLUSTRATIONS

FIGURE 1 : The trial system 4

FIGURE 2 : Second stage punishments of the new CIS 5

FIGURE 3 : The orange and blue curves represent V_n^{Tr} and V_n^{SB} respectively, as a function of ρ 24

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. 26

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. 29

FIGURE 6 : Shift from $r_1(m_1')$ to $r_1(\overline{c})$ 36

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 38

CHAPTER 1 : If we Confess our Sins

1.1. Introduction

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

The traditional solution for this problem is a "trial system". In a trial system, if the principal suspects the crime has been committed, she initiates an investigation aimed at obtaining evidence. Based on the evidence, the principal forms beliefs about the guilt of each agent and chooses punishments accordingly. Only agents whose evidence strongly indicates guilt are punished - agents are punished if they are found to be guilty beyond "reasonable doubt". The merit of this system is that the evidence is more likely to point to guilt if the agent is indeed guilty than if he is not.

In this paper, however, I argue that trial systems are not optimal. There are other systems which generate a larger social welfare, which will be understood as a weighted average between society's desire to punish those who are guilty and to protect those who are innocent. In particular, the optimal system will be a "confession inducing system" (CIS). A CIS has two stages. In the first stage, before the investigation begins, all agents are given the opportunity to confess the crime, in exchange for a guaranteed punishment

independent of any evidence which might be gathered in the future. In the second stage, if necessary, the principal conducts an investigation, and, based on the information gathered, chooses the punishments, if any, to apply to agents who chose not to confess in the first stage. It essentially is a trial system 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 behind self-reporting is that firms which infringe environmental regulations are able to contact the corresponding law enforcement authority and self-report this infringement in exchange for a smaller punishment than the one they would have received if they were later found guilty. Another 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.

These type of systems have received quite a lot of attention in the literature on the economics of crime and law enforcement, which has highlighted some of its advantages.[1] This paper contributes to this literature in two ways. First, in its approach. Unlike most of the literature, which performs pairwise comparisons between the trial system and an alternative system (like plea bargaining or self-reporting), I use some of the techniques from mechanism design to find the optimal system.[2] I believe this is an important contribution in that it makes unnecessary the pursuit of a better system, at least in the context of my model. Second, I highlight advantages of these systems which have not been yet been accounted for. There are two main arguments in favor of CIS's that are prevalent in the literature. First, they require less resources - for example, with plea bargaining, which is how more than 97% of all criminal cases in the United States are resolved (Dervan and Edkins (2013)), it is not necessary to pay all the lawyers, judges and jurors one would have to pay otherwise.[3] And second, they reduce risk - for example the risk of seeing those who are guilty escape unpunished (Grossman and Katz (1983)). I argue that, even if there are no costs and even

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

[2] In an independent work, Siegel and Strulovici (2015) follow a similar approach, which I discuss in more detail in the related literature section.

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

if everyone is risk neutral, there are still two advantages to CIS's. First, in a CIS, the principal is able to threaten agents who refuse to confess with a harsher punishment than they would have received in a trial system in the event of a conviction. Second, and most importantly, CIS's explore the correlation between the agents' innocence in that, when an agent confesses to be guilty, he is also indirectly providing the principal with information relative to other agents. The following example illustrates these two advantages.

Imagine that, in a small town, there has been a fire which damaged a local forest. The principal suspects that it might not have been an accident. She has done some investigative work and has narrowed 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 requests that a modern device be sent to her from a different country, which will allow for the analysis of the residues collected from the forest and will shed light on what has happened.

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 if agent 1 is guilty. 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 the agent 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 the agent 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 waits for the new device to arrive, collects and analyzes the residues (i.e. observes θ_1), forms beliefs about the guilt of the agent and then chooses 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 concerned the principal is about wrongly punishing innocent agents.

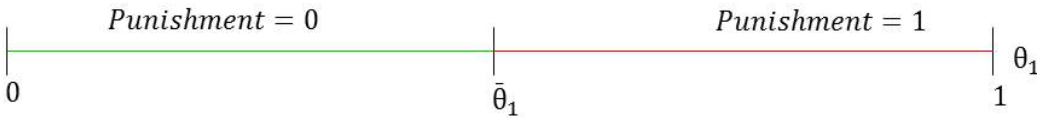

Figure 1: The trial system

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 can commit to punishments, and consider the following alternative. Imagine that, before the new device arrives, the principal approaches the agent and gives him the opportunity to confess in exchange for a 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 $\frac{3}{4}$ is chosen exactly to make the agent indifferent when guilty, giving him just enough incentives to confess, while, if innocent he prefers not to. Therefore, in this alternative CIS, the agent'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 CIS's with respect to trial systems on the grounds that the latter uses less resources - if the agent confesses the crime there is no need to collect evidence. In this paper, because I assume there are no costs of any nature, these two systems are considered equivalent.

I now show that it is further possible to create a new CIS which reduces the expected punishment of an innocent agent (reduces B_1^i) while keeping the guilty agent's punishment constant ($B_1^g = \frac{3}{4}$). I 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 change is that, when

the agent is guilty, he no longer prefers to confess. So, one must increase the second stage punishment, in order to provide him with just 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.

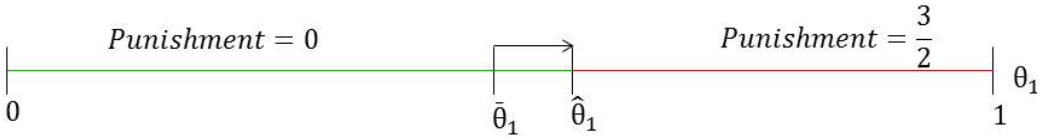

Figure 2: Second stage punishments of the new CIS

In this new CIS, $B_n^g = \frac{3}{4}$ because the agent is confessing the crime when guilty, but $B_n^i = \frac{3}{8} < \frac{1}{2}$, i.e. 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 is 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 administered in the trial system. However, it seems questionable to me whether it is desirable or even possible to construct a system which 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 which inflicts a punishment of, say, 50 years of imprisonment or worse on 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 CIS's that are strictly preferred to any trial system? In general, the answer is *yes*, provided there are at least two agents.

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 collects all the available evidence, and then chooses how to punish all agents. Consider the following alternative. Before initiating the investigation, the principal gives agent 1 the opportunity to confess in exchange for a constant punishment which, if guilty, leaves him indifferent between confessing and refusing to confess. After agent 1 has chosen to either confess or not, the principal initiates an investigation aimed at producing evidence, which is used to select the punishments of all other agents (as well as agent 1 if he has chosen 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 have collected the same evidence as under 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. Therefore, the decision the principal makes with respect to the other agents is more accurate, as she has more relevant information. For example, imagine that, by the nature of the crime, the principal believes that there is at most one guilty person. In this case, a confession by agent 1 leaves the principal very certain that the other agents are innocent, and so she is less likely to make the mistake of punishing them. In other words, there are informational externalities to 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 alternative - in the optimal CIS every agent is given the opportunity to confess - it illustrates the shortcomings of the trial system and highlights the informational 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, which are likely to be committed by an organized group - for example, in anti-collusion legislation, because 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 the example illustrates is that the same argument can 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 be informative with respect to the innocence of others. While for most crimes the guilt of the agents is negatively correlated (so that, if the principal knows that one of the agents is guilty, she is more likely to believe that the other agents are innocent) it is also easy to think of crimes where there is positive correlation. For example, a firm's confession of having reported a smaller profit when filling out the previous year's tax returns, in order to pay less taxes, might inform the principal that other firms in the same sector might have followed a similar practice. Therefore, I believe this informational argument is quite broad and applies to most crimes. Notice also that, for such an argument to follow, it is necessary that there are multiple agents - multiple people who could have conceivably committed the crime. This is always going to be the case if the opportunity to confess is given early enough in the criminal process, when everyone is a suspect, which is something that a system like self-reporting accomplishes by granting the initiative to confess to the agents. By contrast, in plea bargaining, it is the prosecutor who, further along in the criminal process, approaches the generally single agent to seek a confession, which negates this informational advantage, as there cannot be information externalities if one is considering a single agent.

In the first part of the paper, I conduct my analysis under the assumption that the principal has commitment power. In the optimal CIS, the principal uses her commitment power to i) impose small punishments on knowingly guilty agents (the ones who confess), and ii) punish knowingly innocent agents (the ones who refuse to confess). Assumption ii) is

particularly problematic to me in that it causes a disconnection between what the principal would prefer to do at the trial stage and what the mechanism asks her to do. In particular, simply by observing the agent has chosen not to confess, the principal is able to infer he is likely to be innocent. And yet, the mechanism requires the principal to ignore this belief and punish the agent if the evidence is sufficiently high. Because of this, in the second part of the paper, I consider the principal's problem when she has limited commitment power. I consider two cases.

First, I consider the class of renegotiation proof mechanisms, where only i) is permitted. I call these mechanisms renegotiation proof because if the principal is supposed to punish an agent she knows is innocent, both her and the agent would have an incentive to renegotiate such punishment, as they would both prefer a smaller one. In this setup, I show that CIS's are still optimal - they are preferred to any other renegotiation proof mechanism. However, this new CIS is markedly different than the one with commitment power in that it no longer completely separates guilty from innocent agents. If only innocent agents refused to confess, then the principal would know they were innocent and would choose to acquit them. But, in that case, there would be no reason for guilty agents to confess. Hence, the optimal renegotiation proof CIS will be semi-separating: a fraction of guilty agents refuses to confess so that it is possible that either innocent or guilty agents end up convicted at the second stage trial. Nevertheless, even though the principal has less commitment power, it is still the case that this CIS is strictly preferred to the trial system, which shows that the superiority of CIS's with respect to trial systems does not depend on assumption ii).

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 as confessions are no longer sustainable because an agent who is revealed to be guilty is shown no leniency.

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 which 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 which 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 CIS's 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 which 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.

1.2. Related Literature

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

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

In an independent work, Siegel and Strulovici (2015) consider a setting with a risk averse principal and a single risk averse agent and analyze alternatives to the traditional criminal trial procedure, where agents are either convicted or acquitted. The authors demonstrate

that there is a welfare gain in increasing the number of verdicts an agent can receive: so, for example, a verdict of "not proven" in addition to the traditional verdicts of "guilty" and "not guilty". The paper also considers plea bargaining, interpreting a guilty plea as a special type of a third verdict that agents can choose, and show it is uniquely optimal in such a setup.

The main difference between these papers and mine is that the argument I make about the optimality of CIS's does not depend on the agents or the principal being risk averse (as, at least in main text, these are assumed to be risk neutral) nor on them being cheaper (as there are no costs in my paper), but, rather on the fact that CIS's a) explore the correlation between the agents' innocence and b) provide the opportunity for the principal to overpunish in order to induce confessions.[4]

A feature common to these papers is that they have assumed that the law enforcer has commitment power. There have been different articles, particularly in the plea bargaining literature, that have discussed the implications of limiting that commitment power. Baker and Mezzetti (2001) assume that the prosecutors are able to choose how much effort to put into gathering evidence about the crime, after having given the opportunity for the defendant to confess. Given that the prosecutors have no commitment power, in equilibrium, only some guilty agents will choose to confess, while the remaining ones (alongside the innocents) will not. This is because, if all guilty agents confessed, there would be no incentive for the prosecutor to exert any effort, which, in turn, would induce the guilty agents not to confess. This type of equilibrium is a common occurrence when limited commitment power is assumed - see for example Kim (2010), Franzoni (1999) or Bjerk (2007). In section 6, I consider the implications of reducing the principal's commitment power and find that the

[4]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.

optimal mechanism has this same feature: in equilibrium a fraction of guilty agents prefers not to confess.

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 which 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 which also considers multiple agents - see Spagnolo (2006) for a literature review.

In terms of the methodology, the environment studied in this paper is characterized by the fact that there is a single type of good denominated "punishment". The allocation of that good has implications not only to the agents but also to the principal's expected utility. There is some literature on mechanism design which considers similar environments by assuming that the principal cannot rely on transfer payments. In these environments, because the principal is deprived of an important instrument in satisfying incentive compatibility, it is necessary to find other ways of screening the different types of agents. One such way is to create hurdles in the mechanism that only some types are willing to go through. For example, Banerjee (1997), in solving the government's problem of assigning a number of goods to a larger number of candidates with private valuations, argues that, if these candidates are wealth constrained, it is efficient to make them go trough "red tape", in order to guarantee that those who value the good the most end up getting it. In Lewis and Sappington (2000), the seller of a productive resource uses the share of the project it keeps in its possession as a tool to screen between high and low skilled operators which are wealth-constrained. Another approach is to assume 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.[5]

1.3. Model

There are N agents and a principal. Each agent n randomly draws a type $t_n \in \{i, g\} \equiv T_n$, which is his private information - each agent n is either innocent (i) or guilty (g) of committing the crime. Let $T = \{T_n\}_{n=1}^N$ 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 that 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 which are likely to have been committed by an organized group of agents (conspiracy crimes). For example, imagine that agents 1 and 2 rob a bank together. It would be very likely that agent 1 would know that both him and agent 2 are guilty of committing the crime. In section 7.2., I extend the model in order to consider this type of information structure and show that the same intuition carries through. In particular, the optimal system can be interpreted as an "extended" CIS, where each agent is given the opportunity to incriminate other agents when confessing.

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 $\Theta = \{\Theta_n\}_{n=1}^N$ denote the set of all possible evidence vectors, while Θ_{-n} denotes the set of all possible evidence vectors that exclude only agent n's evidence level.

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

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

Let $l(\theta_n) = \frac{\pi(\theta_n|t_n=g)}{\pi(\theta_n|t_n=i)}$ be the evidence likelihood ratio. I assume that l is differentiable and strictly increasing. This implies that the larger the realized θ_n is, the more likely it is that agent n is guilty. I also assume that $\lim_{\theta_n \to 0} l(\theta_n) = 0$ 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 that each agent n's utility is given by $u^a(x_n) = -x_n$, where $x_n \in \mathbb{R}_+$ represents the punishment agent n receives - it could be time in prison, community service time, physical punishment or a monetary fine. Each agent simply wants to minimize the punishments inflicted upon him. I make the assumption that agents are risk neutral in order to distinguish my argument from the one, for example, of Grossman and Katz (1983) (which I discuss in the related literature section), where the advantage of CIS's 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 CIS's 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 who receives it is innocent or guilty. I assume that 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, which I normalize to 1. In either case, deviations from the preferred punishment induce a linear cost to the principal.[6] This punishment of 1 that "fits the crime" is exogenous to the model and is likely to be influenced by the nature of the crime - the punishment that fits the crime of murder is larger than the punishment that fits the crime of minor theft. As it will be become clear in section 4, this will be the punishment imposed in the trial system 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 apparently ignore one of the most important goals of any criminal justice system: to deter crime. In particular, if the goal of the principal was 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 which is chosen in an appropriate way.[7]

[6] 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.

[7] See Figure 4 and the subsequent discussion.

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

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

1.4. Trial System

I define the trial system as a system 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 upon the agents, who do not have any active role. Let $X^{Tr} = \left\{x : \Theta \to \mathbb{R}_+^N\right\}$ be the set of all possible allocations which are implementable through a trial system. In the optimal trial system, the principal chooses an allocation from X^{Tr} in order to maximize her ex-ante expected utility, which is given by

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

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

Notice that $V(x) = \sum_{n=1}^{N} V_n(x_n)$ where

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

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

$$\int_{\theta \in \Theta} \left(\pi(t_n = g | \theta) - \alpha \pi(t_n = i | \theta)\right) \pi(\theta) x_n(\theta) \, d\theta - k \tag{1.1}$$

where k is some constant, independent of x_n, $\pi(\theta) = \sum_{t \in T} \pi(t, \theta)$ for all $\theta \in \Theta$ and 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.

Condition (1.1) displays the simple basis for the principal's decision in a trial system. If $\pi(t_n = g|\theta) > \alpha\pi(t_n = i|\theta)$, 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, 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 she convicts the agent if and only if his evidence level θ_n is above such threshold.

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

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 might be correlated and each agent's evidence level is informative of that agent's type.

1.5. 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.[8] 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,

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

where $\pi\left(t,\theta|t_n\right)$ 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.

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

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

This upper bound is meant 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

[8] In one of the extensions, in section 7.4, I consider a different time frame, where the principal only proposes the mechanism after privately observing the evidence.

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 that of selecting an allocation from X^{SB} that maximizes V, subject to (IC) and (UB). As in the previous section, because it is possible to write $V(x) = \sum_{n=1} V_n(x_n)$, 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 that follow an innocent report to equal those which follow a guilty report. This new allocation would be incentive compatible (as it would not depend on the agent's own report) and would strictly increase the expected utility of the principal (because it would strictly decrease the expected punishment of the innocent agent).

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

$$\pi(t_n = g) B_n^g - \alpha \pi(t_n = i) B_n^i - k \tag{1.2}$$

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

The remaining incentive constraint can be written as

$$B_n^g \leq \int_{\theta \in \Theta} \sum_{t_{-n} \in T_{-n}} \frac{\pi\left(g, t_{-n}, \theta\right)}{\pi\left(t_n = g\right)} x_n\left(i, t_{-n}, \theta\right) d\theta \tag{1.3}$$

From (1.2) and (1.3), it follows that it is optimal to set $x_n\left(g, t_{-n}, \theta\right) = B_n^g$ 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 (1.2) and (1.3) 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. Therefore, in an optimal solution, (1.3) must hold with equality. By plugging (1.3) into (1.2), 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 $x_n\left(i, t_{-n}, \theta\right) \in [0, \phi]$ for all $t_{-n} \in T_{-n}$ and $\theta \in \Theta$, in order to maximize

$$\int_{\theta \in \Theta} \sum_{t_{-n} \in T_{-n}} \left(\pi\left(g, t_{-n}, \theta\right) - \pi\left(i, t_{-n}, \theta\right)\right) x_n\left(i, t_{-n}, \theta\right) d\theta - k \tag{1.4}$$

subject to

$$\int_{\theta \in \Theta} \sum_{t_{-n} \in T_{-n}} \frac{\pi\left(g, t_{-n}, \theta\right)}{\pi\left(t_n = g\right)} x_n\left(i, t_{-n}, \theta\right) d\theta \leq 1 \tag{1.5}$$

Condition (1.5) simply states that B_n^g, which is equal to the left hand side of (1.5) by (1.3), 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. This case is also interesting because, as I show below, when the principal has limited commitment power, she will no longer be able to impose punishments that are larger than 1.

If $\phi = 1$, constraint (1.5) 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 (1.4) that the optimal punishment to be inflicted upon an innocent agent is 1 if

$$\pi\left(t_n = g | t_{-n}, \theta\right) > \alpha\pi\left(t_n = i | t_{-n}, \theta\right) \tag{1.6}$$

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 (1.6). The optimal allocation is implemented by having guilty agents confess and innocent agents refusing to.

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

CIS's 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 CIS's already exist under a variety of forms, like self-reporting regulation in environmental law and plea bargaining in criminal law.

Recall that, in a trial system, an agent has no other choice but to go to trial and be punished if $\pi(t_n = g|\theta) > \alpha\pi(t_n = i|\theta)$, 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 $\pi(t_n = g|t_{-n}, \theta) > \alpha\pi(t_n = i|t_{-n}, \theta)$. 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. For example, in a case where the principal is convinced that there is at most one guilty agent, observing a confession informs the principal that all other agents are very likely to be 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 refusing 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 $N = 2$ and that the prior distribution of guilt is symmetric and given by the following table:*

Table 1: Prior Distribution of Guilt

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

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

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

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

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

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 decision of the other agent. In particular, it will be the case that, if the other agent chooses to go to trial, 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 confesses, 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 is innocent and chooses to go to trial, 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 is guilty and chooses to confess. 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*.*

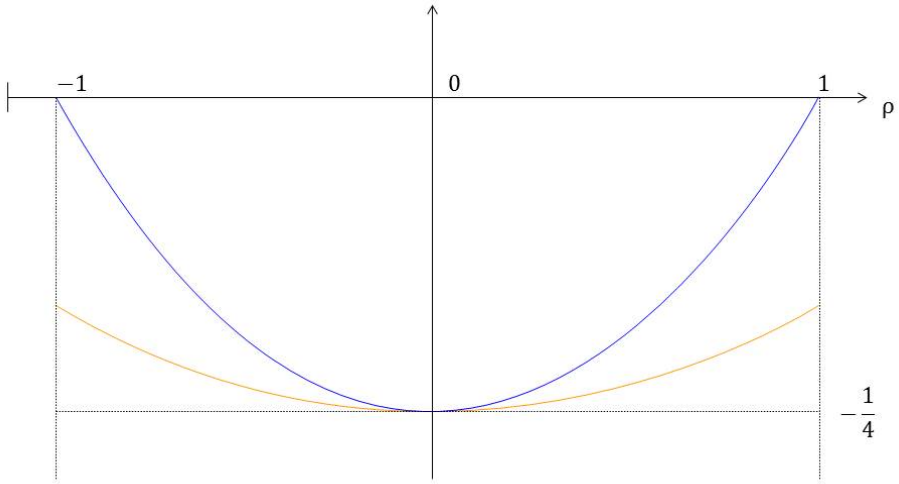

Figure 3: The orange and blue curves represent V_n^{Tr} and V_n^{SB} respectively, as a function of ρ

There is one interesting property of any optimal allocation that I believe is worth emphasizing. Notice that condition (1.6) 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 she 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 that he is guilty) will not be used against him when determining what punishment he is to receive. So, in a way, the optimal allocation is in contrast with the American criminal law practice of the *Miranda warnings*, or, at least, with the part where an agent is told that everything he says might be used against him in court. According to this analysis, the principal should be doing the exact opposite: she should be providing a guarantee that she will **not** use this information against the agent, which, 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, which I study 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.[9] 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

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

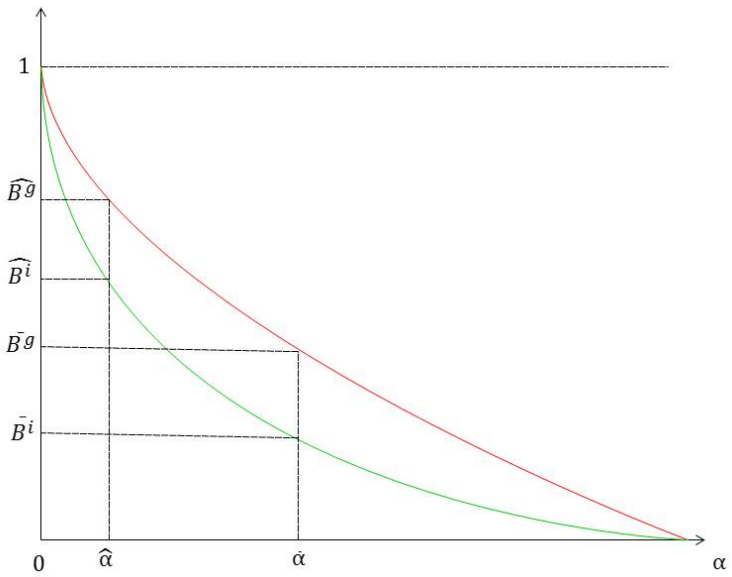

Figure 4: The green and red lines represent the expected punishment of a given agent when innocent and guilty respectively as a function of α.

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.

Notice that Figure 4 identifies the set of "second-best efficient" points: for each expected punishment of 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, she 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 $\alpha = \overline{\alpha}$.[10]

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 measuring 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.

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

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

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

[10] Another way to see this is by realizing that, when $\phi = 1$ and for each agent n, the objective function of the principal can be written as

$$\pi(t_n = g) B_n^g - \alpha \pi(t_n = i) B_n^i$$

and so, if $\alpha = \overline{\alpha}$, where

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

it becomes proportional to $\left\{B_n^g - B_n^i\right\}$.

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.

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 with respect to the case of $\phi = 1$. First, if an 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 (1.5). Hence, if $\phi = 1$ then $\lambda_n = 0$. But if ϕ is sufficiently large (bigger than $\overline{\phi}_n > 1$, which 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.[11]

Figure 5 depicts the evolution of the solution as ϕ increases for some arbitrary agent.

If ϕ is close to 1 - in Figure 5, if $\phi \leq \overline{\phi}_n$ - constraint (1.5) 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. 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

[11] Recall that the simplified problem does not depend on guilty punishments. The only requirement is that the expected punishment for the guilty agent is 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.

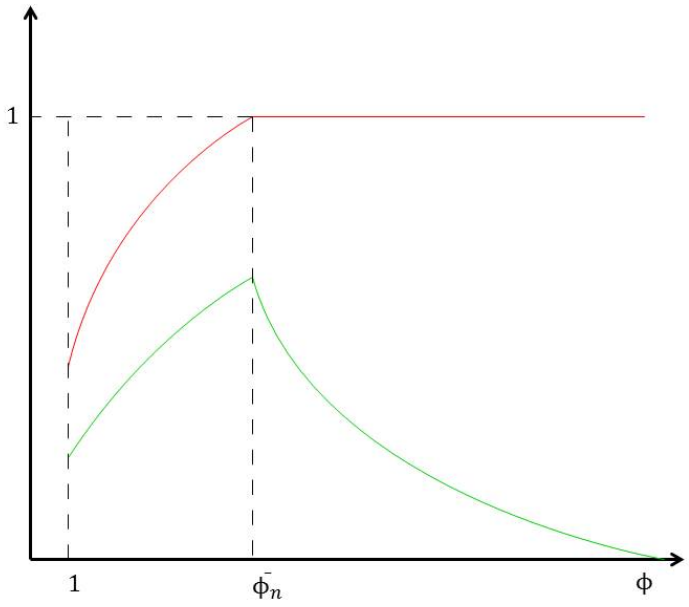

Figure 5: Evolution of the agent's expected punishment as a function of ϕ. The red and green curves represent the expected punishment when the agent is guilty and innocent respectively.

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}.

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 each agent's type is also correlated with the evidence. The idea is that the principal can simply punish in 1 all guilty reports, and set a sequence of punishments 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 setting a very high (close to 1) standard of proof at the trial stage, but imposing arbitrarily large punishments in the event of a conviction. The result follows because guilty agents are infinitely more likely than innocent agents to generate an evidence level that is close to 1.

The problem of excessive commitment power

The CIS which 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.[12] 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

[12]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.

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). 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.

1.6. Limited Commitment Power

In this section, I analyze the problem the principal faces in 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, 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 a message m_n from the message set M_n, prior to knowing the evidence. Let $M = M_1 \times ... \times M_N$ and refer to m as a generic element of M. I also give the usual interpretation to m_{-n} and

M_{-n}.

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

I call each profile (x, σ) a system and evaluate it according to the principal's expected utility. In particular, I denote by $\widehat{V}(x, \sigma)$ the principal's expected utility of pair (x, σ), where

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

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

$$\begin{aligned} &- \int_{\theta \in \Theta} \int_{m_{-n} \in M_{-n}} \pi^{\sigma}(m_{-n}, \theta | t_n) \, x_n(m_n, m_{-n}, \theta) \, dm_{-n} d\theta \\ &\geq - \int_{\theta \in \Theta} \int_{m_{-n} \in M_{-n}} \pi^{\sigma}(m_{-n}, \theta | t_n) \, x_n\left(m_n', m_{-n}, \theta\right) dm_{-n} d\theta \text{ for all } m_n' \in M_n \end{aligned} \tag{1.7}$$

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 (1.7) holds, I say that the system (x, σ) is incentive compatible.

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

Definition 1 *A CIS (x, σ) is such that, for all n,*

i) In equilibrium, at most two messages are sent with positive probability by each agent: a confessing message c and a non-confessing message $\overline{c}$.

ii) If an agent confesses (sends message c) 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.

1.6.1. 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 an arbitrary 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.[13]

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

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

Definition 2 *The system (x, σ) is renegotiation proof if and only if, for all n, m and θ,*

$$x_n(m, \theta) \leq \gamma_n^{\sigma}(m, \theta) \tag{1.8}$$

If system (x, σ) is renegotiation proof, then I say that mechanism x is renegotiation proof.

The first thing to notice is that condition (1.8) effectively imposes an upper bound of 1 to all punishments, which prevents the principal from overpunishing. In particular, proposition 5 no longer holds as the characterization of the optimal system is independent of ϕ, provided it is (weakly) larger than 1. Notice also that the CIS described in the previous

[13] If there are multiple maximizers, $\gamma_n^{\sigma}(m, \theta)$ is defined to be equal to the smallest one.

section, which 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 1, which delimits the message set of each agent.

Lemma 1 *Without loss of generality, it is possible to set* $M_n = \mathbb{R}_+ \cup \{c\}$ *for all* n.

Proof. See appendix. ■

The meaning any message conveys is given by the belief the principal forms when she receives it. In Lemma 1, 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 2, 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 6, I show that $\left(x^{\sigma^{RP}}, \sigma^{RP}\right)$ 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 2 *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

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

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, which leaves him indifferent to sending any other message if he is guilty. 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 her preferred punishment conditional on what she has learned in the first stage, and on the evidence, and so $x_n^\sigma(m_n^\sigma, m_{-n}, \theta) = \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta)$. 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 as close to optimal as possible, which implies that $x_n^\sigma(m_n, m_{-n}, \theta) = \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta)$ 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 6 *A CIS is optimal within the set of incentive compatible and renegotiation proof systems.*

Proof. See appendix. ■

In proposition 6, I show that it is optimal for agents to send at most two messages: the confessing message c and a non confessing message $\overline{c}$. The argument is as follows. Take any strategy profile σ and label message m_n^σ as $\overline{c}$. Suppose that, without loss of generality, agent

1 is sending a second non-confessing message m_1' in addition to message $\overline{c}$. As mentioned above, each message m_1 is identified by its "guiltiness" ratio $\frac{\sigma_1(g,m_1)}{\sigma_1(i,m_1)} \equiv r_1(m_1)$. Suppose that $r_1(\overline{c}) < r_1(m_1') < \infty$. 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 $\frac{\sigma_1(g,m_1')-v}{\sigma_1(i,m_1')} = r_1(\overline{c})$, it is possible to increase the expected utility of the principal (see Figure 6).

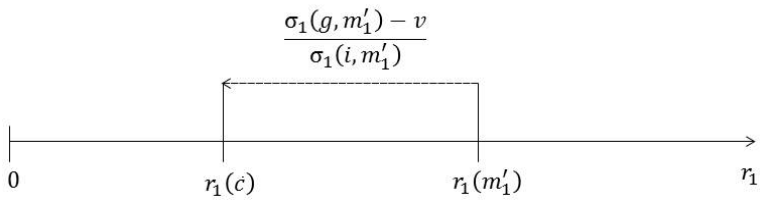

Figure 6: Shift from $r_1(m_1')$ to $r_1(\overline{c})$

The expected punishment of agent 1 is unchanged regardless of whether he is innocent or guilty, because message $\overline{c}$ is still available and the expected punishment of sending it remains the same (given that ratio $r_1(\overline{c})$ 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 6 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. She was only able do this because she was able to commit, which would mean having a set of laws and regulations for judges, lawyers and jurors to follow, which would not necessarily be designed to assess the agents' guilt. But under this new CIS this is no longer necessary. 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 7, I show that such a system is not optimal.

Proposition 7 *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 7 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 one.

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 $\tau_n \in [0,1]$. Consider the punishments of agents that choose not to confess. If the other agent does not confess the crime (chooses to play $\overline{c}$), agent n is punished if and only if*

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

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

and only if

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

Notice that if $\tau_1 = \tau_2 = 0$ *this CIS becomes the trial system in that no agent confesses, and threshold* $\theta_n^{RP}(\overline{c}, \theta_{-n})$ *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* $\tau_n = 0$ *and* $\tau_{-n} = 1$*. This means that either* $\theta_1^{RP}(\overline{c}, \theta_{-n}) \neq \theta_1^{SB}(t_{-n} = i)$ *or* $\theta_2^{RP}(\overline{c}, \theta_{-n}) \neq \theta_2^{SB}(t_{-n} = i)$ *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,

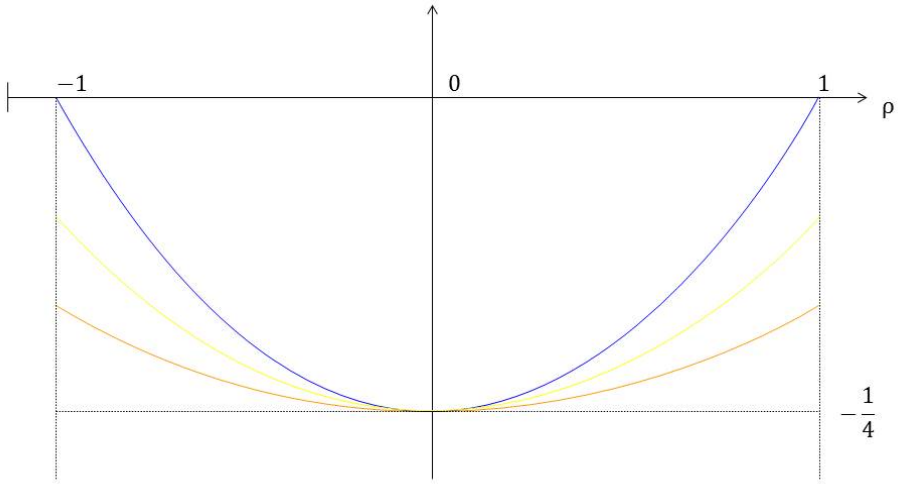

Figure 7: The orange, yellow and blue curves represent V_n^{Tr}, V_n^{RP} and V_n^{SB} respectively, as a function of ρ

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 to each confession.

1.6.2. Sequentially Optimal Mechanisms

CIS's are based on the assumption that the principal is able to partially forgive a guilty agent who confesses, precisely in order for him to confess. However, knowing only guilty agents confess, it is not ex-post optimal for the principal to show leniency towards them. Hence, if the principal does not have commitment power, she will be unable to implement such confession inducing mechanisms. In this section, I analyze what mechanism should the principal implement if 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.

Definition 3 *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 her ability to collect any information from the agents. Imagine that agent n is sending two distinct messages a and b. For these messages to convey any information to the principal, it must be that they are sent with different probabilities by the innocent and the guilty types. Suppose a is more likely to have been sent by the innocent type than b. Knowing this, the principal has no choice but to be more lenient towards agents that have sent message a. But then, no agent would ever send message b. It follows that, if the principal is unable to recover any information from the agents, all we are left with is the trial system.

Proposition 8 *If the principal has no commitment power, a trial system is optimal.*

1.6.3. How much commitment power does the principal have?

This paper characterizes the principal's preferred mechanism under three different assumptions regarding her commitment power: full commitment power, no commitment power and

an in-between assumption, where the principal is only unable to commit not to renegotiate. But which of three assumptions is more reasonable?

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

However, looking at contemporaneous societies one can see that there are several examples where leniency is shown towards agents who confess to have committed a crime. The method modern societies seem to follow, in order to commit to such leniency, is to use law. For example, plea bargain deals are protected under Rule 11 of the Federal Rules of Criminal Procedure, which ensures the prosecutor cannot go back on his word once he has obtained the confession from the agent. But if societies can use law as a commitment device, one could argue that the relevant analysis should be the one that assumes full commitment power by the principal. The problem with this argument has to do with the human element that is present in judging an agent's guilt. Consider the optimal 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 chose not to testify. As Laudan (2006) points out, this practice precludes important information from the trial and conditions how jury members assess the defendant's guilt. Whether these recommendations are indeed taken into account by the jurors is a matter of discussion: Laudan (2006) cites Posner (1999) on this matter: "Judges who want jurors to take seriously the principle that guilt should not be inferred from a refusal to waive the privilege against self-incrimination will have to come up with a credible explanation for why an innocent person might fear the consequences of testifying".

In my opinion, the proper assumption over the principal's commitment power depends very much on how one feels about these attempts at conditioning guilt assessment. If one believes that police, judges and jurors always follow the law and enforce punishments they know are unfair, then the relevant assumption should be of full commitment power 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 the second stage trial veredict only reflects the jurors true assessment of guilt under x^{RP}.

1.7. Extensions

The main purpose of this paper is to highlight how CIS's are able to explore the correlation between the agents' innocence in order to provide a more efficient alternative to trial

systems. 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 CIS's 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 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, consequently, 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 who confess are also requested to report what they know about the crime.

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 who are guilty, who choose to confess, and those who 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 a perfect separation between those who are guilty and those who 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[14].

1.7.1. Risk Averse Agents

One of the assumptions of this paper is that agents are risk neutral. This might lead the reader to inquire on whether CIS's would still be appealing if agents were risk averse. The concern might be that agents choose to confess because they are risk averse and not because they are guilty. In order to address this issue, in this section, I extend the analysis to consider arbitrary levels of risk aversion for the agents and for the principal in a setup close to that of the independent work of Siegel and Strulovici (2015). Proposition 9 corroborates that paper's result in arguing that enlarging the set of possible verdicts increases the expected utility of the principal, while Proposition 10 can be understood as a special case of their analysis where, given a specific functional form for the agents' utility function, I show that a CIS is still an optimal system.

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

Let $\widetilde{x}^{Tr}$ denote the optimal allocation that can be implemented by a trial system.

Proposition 9 *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 $\lim_{\theta_n \to 1} \widetilde{x}_n^{Tr}((\theta_n, \theta_{-n})) = 1$.*

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

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.

Let $\widetilde{x}^{SB}$ denote the second best allocation.

Proposition 10 *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, she would want to maximize his expected utility. Assuming that $u_n^p(i, \cdot)$ is proportional to $u^i(\cdot)$ implies precisely that - the principal has the same preferences of the innocent agent when she knows him to be innocent. This assumption is convenient in that it guarantees that innocents' incentive constraints do not bind.

Proposition 10 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 punishments could be reduced to 1, which would increase the principal's expected utility and give more incentives for guilty agents to report truthfully). Suppose that, in the optimal allocation,

a guilty agent receives a lottery of distinct punishments. Because the guilty agent is risk averse, he would be willing to accept, as an alternative, a constant punishment larger than the expected punishment of the lottery. The principal would strictly prefer this alternative, as long as she is (weakly) risk averse.

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

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

Proposition 11 *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 she faces a trade-off. On the one hand, she would like to select small punishments in order to spare the innocents as much as possible. But on the other hand, those punishments determine the punishment that guilty agents receive in equilibrium. So, the principal wants to construct a lottery of punishments that is very appealing for those who are innocent but very unappealing for those who are guilty. If innocent agents are more risk averse than guilty agents, then smoothing punishments is relatively better for them, which is why, if $\omega_i > \omega_j$, the punishments innocent agents receive are strictly increasing and continuous until hitting the upper bound of ϕ. If, on the contrary, guilty agents are more risk averse, following a similar strategy would be relatively better to guilty agents. Therefore, even though agents are strictly risk averse regardless of whether they are innocent or guilty, if $\omega_i \leq \omega_j$, it is still better for the principal to impose a risky lottery of punishments, where agents are punished very harshly only for very high levels of evidence, and are acquitted otherwise.

1.7.2. 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 have implicitly ruled out criminal conspiracies. When a group of agents commits a crime together, it seems reasonable to expect them to know the identity of the remaining conspirators. For example, if a group of 3 agents robs a bank, it is very likely that each of them will know the identity of the others. In this section, I extend the model to accommodate for this possibility and investigate how the optimal mechanism changes if the principal believes that a criminal conspiracy might be behind the crime.

I assume that, for each event $t \in T$, there is a commonly known probability $p(t) \in [0,1]$ that each guilty agent knows the identity of the remaining criminals (and so knows vector t). So, for example, if $N = 3$ and $p((g,g,i)) = 0.75$, it means that, when the crime is committed by agents 1 and 2, there is a 75% chance that the agents committed the crime

together and know each other's identity. Hence, in that case, agents 1 and 2 would know that vector (g, g, i) had been realized. With 25% chance, agents 1 and 2 committed the crime independently and do not know whether any of the other agents is also guilty. In either case, agent 3 is innocent and forms beliefs about agents 1 and 2's guilt as before.

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

For simplicity, I consider only the case of $\phi = 1$ and assume that the principal has commitment power.

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

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

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

where $\widehat{\varphi}_n$ is characterized in the proof of Proposition 12.

Proposition 12 *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 her 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 in the standard CIS, agents are given the opportunity to confess. However, they are also asked to report any other information they might have, in particular, whether there are other guilty agents and their identity. By construction of $\widehat{x}^{SB}$, guilty agents are indifferent between confessing and going to trial, while innocent agents refuse to confess. 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 who confess no longer receive a constant punishment. With this information structure, guilty agents might have different beliefs about the guilt or innocence of other agents. This means that a constant punishment which leaves a guilty agent of extended type $\widehat{g}$ indifferent might not leave him indifferent if he has some other extended type. However, these different extended types of guilty agents all have the same beliefs with respect to the evidence the agent himself generates. Therefore, the punishment an agent receives when he confesses only depends on the type of information that other agents grant the principal ($\widehat{t}_{-n}$) and not on the evidence.

I illustrate how this extended CIS works by continuing the previous example.

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 commit the crime, while it is certainly possible that they 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 original assumption that 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 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*

$$x_1((g,g),i,\theta) = \begin{cases} 1 \text{ if } \theta_2 < d_1 \\ 0 \text{ otherwise} \end{cases}$$

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.

1.7.3. Heterogeneous agents

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

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

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

Proposition 13 below characterizes how each agent acts in the optimal CIS when $\phi = 1$ and the principal has commitment power.

Proposition 13 *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 $\beta_n \in [0,1]$,*

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

where $s_n(t_n, \beta_n) \in \{c, \overline{c}\}$ represents the action that agent n with type t_n and 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 larger incentive to confess (and select action c) than those with a smaller 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 is small enough that even if the guilty agent draws $\beta_n = 0$, he still prefers to confess. But establishing such a small punishment leads to innocent agents confessing. For example, if there is no correlation between the agents' 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 who confess (the following argument could also be made if only a small fraction of innocent agents confesses). Of these, only a small fraction is made indifferent (which has a 0 measure) - the pair (g, β_n^g) for each agent n. This means that anytime a guilty agent draws $\beta_n > \beta_n^g$ and chooses to confess, he is strictly better off than choosing not to. Thus, a more successful mechanism would be to punish agents that confess as if

they did not. The principal would still solicit a report from the agents on whether they are innocent or guilty, and punishments that follow an innocent report would still be the same as in the optimal CIS. The difference would be that agents who confess would also face the same lottery of punishments as if they chose not to. They would still have enough incentives to confess (because they would be indifferent), but their expected punishment would be larger.

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

1.7.4. Informed Principal

I consider the same setup as in section 5 but now assume 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 $y_\theta : T \times \Theta \to [0,1]^N$, which 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 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 14 *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 larger 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 14 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 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 15 *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 $y_{\widehat{\theta}}^{IP}(t, \theta) = x^{IP}(t, \theta)$ 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 CIS's are no longer optimal.

1.8. Concluding remarks

The main purpose of this paper is to argue for the virtues of CIS's. The idea that I explore in the paper is that there are information externalities to each confession: when an agent confesses to be guilty he is providing the principal with the information that other agents are likely to be innocent. It then follows from my analysis that all members of the community should be allowed to confess the crime in exchange for a constant punishment, even before any investigation has been initiated. Even though this might appear as a radical suggestion, there are variants of CIS's already in American law. Self-reporting in environmental law works in very much the same way, even though it is mostly motivated

by an attempt to reduce monitoring costs. In that context, agents are firms which are able to confess to have broken environmental regulations in exchange for smaller punishments. And even in criminal law, plea bargaining also allows agents to confess. In this case, agents are defendants and, typically, the bargaining occurs only when there is a single defendant, which largely defeats the purpose of having agents confessing, according to my analysis. A confession produces no information externalities if there are no other agents to consider. In that sense, this paper can be seen as providing an argument for plea discussion to occur earlier in the criminal process, at a time when there are several suspects of committing the crime.

There are, however, a few problems with expanding the policy of self-reporting to criminal cases which are not directly studied in the text. One such problem is that innocent agents might be given enough incentives by guilty agents to confess. 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 in criminal law would then depend on whether it is possible to resolve these type of problems in a satisfying manner. A way to, at least, mitigate them would be to "validate" the confession of any given agent only if the evidence supports the claim.

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

other problems is essential to be able to convincingly argue for the introduction of this type of system in criminal law.

CHAPTER 2 : Inducing Overconfidence

2.1. Introduction

Overconfidence is a phenomenon which has been widely researched in a variety of subjects including economics. There have been several papers which claim to provide evidence in support of the idea that people are overconfident with respect to their own ability (see, for example, Svenson (1981), Guthrie et al. (2001) and Buehler et al. (1994)). While there have been other papers which have not found any evidence of overconfidence (see Clark and Friesen (2009) and Moore and Healy (2008)), there is, at the very least, a justifiable suspicion that, in various settings, people are overconfident.

I believe the phenomenon of overconfidence is interesting because of the puzzle it presents. It is no surprise that people are confused about their ability. After all, ability is unobservable, so each person is limited to observing signals related to it. What is puzzling is the more or less systematic bias toward overconfidence. Furthermore, it is important to understand why overconfidence arises, as it may have serious consequences in people's lives. Overconfidence can be good - there is some evidence that confidence in one's ability improves performance (see Taylor and Brown (1988)). It can also be tremendously bad. People who overestimate their ability will likely make poor decisions - for example, overconfident managers may be too willing to invest in risky projects (Malmendier and Tate (2005)); overconfident entrepreneurs will start businesses they should not have (Camerer and Lovallo (1999)); even overconfident researchers will immerse themselves into solving unsolvable mysteries.

Several theories have been presented to try to explain this phenomenon. In the majority of them, a person becomes overconfident only because of her own decisions.[1] In this paper, I examine the link between overconfidence and the actions that other people take. There is a vast literature, mainly in applied psychology, that has documented external influences on people's expectations - Glasgow et al. (1997) and Smith and Powell (1990), among many others, highlight the impact of parents on their children's expectations; Meyer and

[1] Bénabou and Tirole (2003) is one notable exception, which I discuss below.

Gellatly (1988) argue that setting external goals increases students' self-confidence; Hattie and Timperley (2007) discuss the influence that feedback provided by teachers or parents may have on one's confidence; and Klassen (2004) argues that the actions of one's peers may have a considerable impact on self-confidence. The traditional explanation for the influence of others' actions on self-confidence relies, at least to some extent, on the assumption that agents interpret the information they collect at face value. For example, imagine that some agent's father decides to take over his son's science project and the son ends up receiving an "A" for it. If the son is "naïve" he might become overconfident about his skill simply because he does not take into account that most of the work was performed by his father. While arguments of this nature might sometimes be reasonable, one has to wonder whether, if this type of interaction happened often, the son would eventually realize that the only reason he was getting "A"'s was because of the help he was receiving and not because of his skill. I provide a theory of induced overconfidence where the agent properly discounts any outside influence on his performance.

As an example, consider the relationship between a graduate student and his academic adviser. Upon arrival at the university, the graduate student usually has to go through a series of tests and examinations during his first year. Once his scores are revealed, both he and his adviser use them to update their beliefs about the student's ability. It seems reasonable to assume the adviser will be more willing to provide guidance and advice to the student if she believes the student to be of high ability. Therefore, a student who has performed well in his first year will receive a substantial amount of help, unlike the student who has not.

The key to my argument is the way I model the impact of the adviser's help on the signals that arrive after the first year. Certainly, more help means that the scores the student obtains in future assessments will probably be higher. However, the downside of receiving help would be that these future signals would also be less informative. If the student performs a task on his own and has a good score, he must think that his ability is high. Likewise, if his score is low, he will think his ability is low. However, when the

student receives substantial help, he will be unsure whether the score he obtained was a product of his own ability or of the help he received. Therefore, he will not be as accurate in inferring his ability from the score he observes.

Take two students who happen to have the same ability but different scores in their first year. The one who received a high score will receive more help and will therefore place relatively little weight on future scores. This means that he will be overconfident because his initial score was high. By contrast, someone who has received a low score during his first year will not receive as much help and, as a consequence, will place considerable weight on future scores. Given that the first score was low, the student is likely to improve. Therefore, he is underconfident immediately after the first year but that underconfidence is likely to disappear after future signals. The same does not happen with the overconfident students, who will remain overconfident for longer. It is this asymmetry that causes overconfidence.

There are a number of papers based on this same insight that, while underconfident people will quickly leave that state, overconfident people will not. Zábojník (2004) argues that the opportunity cost of further investigating one's ability increases with ability and so overconfident agents will likely stop investigating sooner than will underconfident ones. Köszegi (2006) states that overconfident people will sometimes obtain less information out of fear that a bad outcome may make them think their ability is low, which has a direct impact on their well-being. Bénabou and Tirole (2002) show that, with time-inconsistent preferences, an overconfident agent may prefer not to acquire costless information. Brocas and Carrillo (2009) argue that overconfident agents require less evidence to support their investment decisions, which may help explain why so many new businesses fail. What distinguishes this paper is that I employ a similar logic in a setup where there are two agents (a father and a son, an adviser and an advisee, etc.) and where one of them strategically tailors his actions to influence the confidence of the other. Benabou and Tirole (2003) consider a setup similar to this, in that they also consider an external figure who influences the agent's beliefs. However, their model does not generate a systematic bias toward overconfident beliefs, which is a contribution of this paper.

The paper is organized as follows. In Section II, I present the model. In Section III, I discuss the conditions under which the agent becomes overconfident, based on the assumption that the more able the agent believes he is, the more help he receives. In Section IV, I discuss why this assumption is intuitive in various settings. In Section V, I conclude.

2.2. The Model

There are two actors in the model: the agent and the person who provides help to the agent, whom I will call the adviser, following the story in the introduction. The agent is endowed with an ability level y which neither he nor the adviser observes. I assume that $y \sim N\left(0, \sigma_y^2\right)$. There are two periods in the model. In the first period, the agent performs a task for which he receives a score s_1. This score will be a function not only of his ability y but also of some independent random variable ε_1. In particular, I assume that

$$s_1 = y + \varepsilon_1$$

where $\varepsilon_1 \sim N(0,1)$. Score s_1 is publicly available so the adviser is assumed to observe it. In the story in the introduction, the score s_1 represents the grades the graduate student receives during his first year.

At the end of period 1, there will be some interaction between the agent and the adviser, which will result in the agent receiving some level of help $h \geq 0$, which the agent is assumed to observe. In Section 4, I discuss this interaction in more detail; for now, I simply assume that the help the agent receives is a function of s_1. In particular, there is $\Psi : \mathbb{R} \to \mathbb{R}_+$ such that $h = \Psi(s_1)$.

In period 2, the student performs a second task which returns score s_2, where

$$s_2 = y + hx + \varepsilon_2$$

I assume that $x \sim N\left(\mu_x, \sigma_x^2\right)$, $\varepsilon_2 \sim N(0,1)$ and that x and ε_2 are independent random

variables. This second score still has the same two components as the first one (i.e., the agent's ability and "noise") but there is a third element, which depends on the help the agent receives. The random variable x is interpreted as the ability of the adviser. The idea is that the help the agent receives determines whether the score depends more on the agent's ability or on the adviser's ability. If h is small, then s_2 will depend mostly on the agent's ability; if h is large, what matters the most is the adviser's ability.

2.3. Induced overconfidence

I am interested in studying the circumstances under which the agent will be overconfident about his own ability y at the end of the second period. Given the structure of the signals the agent observes and by simple Bayesian updating, it follows that

$$y|s_1, s_2 \sim N\left(\widehat{\mu}_y, \widehat{\sigma}_y^2\right)$$

for some $\widehat{\mu}_y$ and $\widehat{\sigma}_y^2$. The issue now is how to define "overconfidence".

One of the most common methods employed by the experimental literature on overconfidence is to ask people whether they believe they are better than the average (or more accurately the median).[2] If more than 50% of the subjects respond "yes", this is taken as evidence of overconfidence, as it is not possible that more than 50% of the people are actually better than the median. In my model, recall that y is symmetric around 0, so there is no distinction between the mean and the median. Furthermore, the agents in my model also reach the end of period 2 believing their own ability is symmetric around $\widehat{\mu}_y$. So, when asked whether they believe their ability is greater than the mean/median, they would say "yes" if $\widehat{\mu}_y \geq 0$ and "no" otherwise, simply because $\widehat{\mu}_y$ represents the conditional expectation, median and mode of their ability.[3]

[2] See Dunning, Heath and Suls (2004) for a review.

[3] For simplicity, I have assumed that ties are solved in favor of the "yes". This assumption has no impact because the event where $\widehat{\mu}_y = 0$ has a 0 measure.

Proposition 16 *If Ψ is increasing, then*

$$\Pr\{\widehat{\mu}_y \geq 0\} \geq \frac{1}{2}$$

The inequality is strict if Ψ is not constant.

Proof. See appendix. ■

Proposition 21 says that, if the help an agent receives is an increasing function of the first score, the probability that the agent believes his ability is greater than the median is larger than 50%. The intuition is the following. Let

$$\widehat{s}_2 = s_2 - h\mu_x$$

so that

$$\widehat{s}_2|y, s_1 \sim N\left(y, Var\left(y|s_1\right) + h^2\sigma_x^2 + 1\right)$$

It follows that the agent's conditional expectation $\widehat{\mu}_y$ is simply a weighted average of s_1 and $\widehat{s}_2$ (and the prior mean 0). However, the weights depend on the help the agent receives. In particular, if help is increasing with s_1, the variance of the second signal $\widehat{s}_2$ is larger, which makes it less informative. As a result, the weight of s_1 is larger when s_1 is larger and smaller when s_1 is smaller. In other words, if an agent draws a large s_1, he is more likely to end up with a conditional expectation close to it than if he draws a small s_1. Notice also that, if Ψ is constant, the weights on s_1 and $\widehat{s}_2$ are also constant and so there is no bias. Furthermore, if Ψ is decreasing, the opposite happens and the probability that an agent's conditional expectation is below 0 is larger than 50%.

Benoit and Dubra (2011) have criticized the previous method of documenting overconfidence. They argue that the fact that most of the population believes that they have a greater ability than the median is evidence of "apparent" overconfidence. The approach taken by those authors is that "real" overconfidence only exists when agents do not have

rational expectations. Given that there are models with agents who have rational expectations where the same result is possible (such as this model), observing that most people believe they have a greater ability than the median is not indicative of overconfidence.

This type of criticism raises the question of whether there is a better, more accurate, way of defining overconfidence. Taken literally, when someone is overconfident, he is too confident with respect to something, in this case his own ability. So, whether an agent is overconfident or not will depend on some comparison between the agent's beliefs and his true ability. Of course, the challenge is to compare a distribution (the agent's beliefs) and a number (the agent's ability). However, in this model, the posterior belief of an agent will be a normal distribution with mean $\widehat{\mu}_y$, which is also the median and mode. Hence, it seems natural to think of $\widehat{\mu}_y$ as the agent's response to the question: "What do you think your ability is?". In this sense, it seems appropriate and intuitive to say that an agent is overconfident if $\widehat{\mu}_y > y$ and underconfident otherwise. Using this definition and maintaining the assumption that the agent indeed has rational expectations does preclude any systematic bias. In particular, if one defines o to be such that

$$o \equiv \widehat{\mu}_y - y$$

it follows that $E(o) = median(o) = 0$. Therefore, if one is of the opinion that overconfidence exists if and only $E(o) > 0$ or if and only $median(o) > 0$, then it must be the case that overconfidence and rational expectations are incompatible, just as it is understood by Benoit and Dubra (2011). However, it is not the case that overconfidence (defined as either $E(o) > 0$ or $median(o) > 0$) is incompatible with Bayes' updating.

The rational expectations assumption can be thought of as three assumptions put together: first, the agent knows the prior distribution of his own ability y; second, the agent knows the distributions of scores s_1 and s_2; third, the agent updates his beliefs about his own ability by Bayes' rule. The first of these assumptions is particularly controversial. The assumption that the agent knows the distribution of his own ability is made mostly out of convenience and not because there is a particularly compelling reason to think that agents

are magically born knowing such a thing. In this context, I believe that an equally (if not more) compelling way to model the agent's behavior is to assume that he has an uninformative prior rather than the correct one. In particular, it seems reasonable to think that, while $y \sim N\left(0, \sigma_y^2\right)$, the agent will believe $y \sim N\left(0, \delta_y^2\right)$, where $\delta_y > \sigma_y$, so that, in a way, the agent is more confused about his ability than he would be if he actually knew its prior distribution. Notice in particular that, as δ_y^2 increases, the prior belief of the agent becomes more and more uniform.

Assuming that the agent's prior belief about y is distributed according to $N\left(0, \delta_y^2\right)$, after having observed scores s_1 and s_2, the agent will believe

$$y|s_1, s_2 \sim N\left(\widetilde{\mu}_y, \widetilde{\sigma}_y^2\right)$$

Finally, let

$$\widetilde{o} = \widetilde{\mu}_y - y$$

so that $\widetilde{o}$ represents the overconfidence of the agent under this new prior.

Proposition 17 *If Ψ is increasing and $\delta_y > \sigma_y$, then*

$$E\left(\widetilde{o}\right) \geq 0 \text{ and } median\left(\widetilde{o}\right) \geq 0$$

Both inequalities are strict if Ψ is not a constant.

Proof. See appendix. ■

By simply removing the assumption that the agent knows the prior distribution of ability and instead assuming that the agent has a prior with a larger variance (and, in the limit, an uninformative prior), it is possible to have overconfidence in the most intuitive of definitions. Not only is this definition of overconfidence intuitive, but there are also examples of empirical papers which document the existence of "expected overconfidence" (where $E\left(\widetilde{o}\right) > 0$). For example, in Smith and Powell (1990), a survey of college seniors

was conducted, in which they were asked their best guess about their own future earnings as well as the earnings of the rest of their cohort. The results were that the average of the answers regarding one's own earnings was higher than the average of the answers regarding the rest of their cohort - a clear indication of expected overconfidence. I believe that it is important to highlight that, even though the assumption of rational expectations must be removed for Proposition 22 to hold, this does not imply the agent is "irrational". He is not endowed with any bias and does Bayes' updating as before. It is simply the case that he does not know about his prior distribution of ability and, as a result, relies less on his prior beliefs and more on the signals he observes.

2.4. Why is help increasing?

In the previous section, I have argued that, if the help an agent receives is an increasing function of the initial score s_1, there will be overconfidence as I have defined it. The purpose of this section is to discuss several circumstances where such an assumption is reasonable and intuitive. I provide three separate models that result in an increasing help function. The common thread in all of them is that a large initial score s_1 leads to a public belief that the agent has high ability. As a result, in the different settings considered, there will be a desire on the part of others to help/cooperate with that agent for their own benefit.

2.4.1. Adviser/Graduate Student

Consider the story from the introduction where a graduate student is completing his Ph.D. The student has some unknown ability y and, after the first year, receives score s_1. At the end of the first year, the student is assigned to an adviser who must choose how much help $h \in \left[0, \overline{h}\right]$ to provide to the student. The help the student receives impacts his future score s_2 as in the previous section.

Suppose the adviser's utility is given by $g(s_2) - ch$, where $g'(s_2) > 0$ for all s_2 and $c > 0$. Assume $\mu_x > 0$ so that the adviser's help, at least on average, does not harm the student, and let $\Psi(s_1)$ denote the optimal help choice made by the adviser.

Proposition 18 *If $g''(s_2) > 0$, then Ψ is increasing.*

Proof. See appendix. ■

By providing help, the adviser incurs a marginal cost c and causes a constant marginal increase of x of the score s_2. However, the adviser is impacted more from the constant increase in the score if the score is large. Thus, providing help to the student gives more benefits to the adviser if she believes the score is likely to be larger, which is why she is more willing to provide help if she believes the student is of high ability.

There are several different reasons why an adviser would have such preferences. Maybe the adviser receives recognition only for her better students, rather than for their mean quality. Or perhaps the adviser benefits only from cooperating with the students perceived to be better, so that there would be an added benefit associated with skilled students. Or it might even be that the adviser believes that highly skilled students are more likely to be in a position of influence in the future, which might benefit the adviser indirectly. As a result, the adviser might be particularly interested in developing a high-ability student's career by providing help in obtaining high scores. What is important for the argument to hold is that the desire to help those students who appear more skilled is greater than the desire to help those who appear less skilled.

2.4.2. Child/Parents

Consider the interaction between a child and his parents. In particular, think of how parents choose how to help their children succeed. Typically, the choice parents make is not so much whether to help their child but rather how to allocate such help.

Imagine that there are N different activities or skills for which an individual is endowed with an ability level y^n, where $n = 1, ...N$. Assume that $y^n \sim^{iid} N\left(0, \sigma_y^2\right)$. In the first period, the child performs N tests for which she receives the corresponding public scores $\{s_1^n\}_{n=1}^N$, where $s_1^n = y^n + \varepsilon_1^n$. As in the previous section, for all n, $\varepsilon_1^n \sim^{iid} N(0,1)$ and is independent of all other random variables.

In the second period, the child performs another set of N tests. However, the parents

are able to help the child in each of these. In particular, for each task n, the score the child receives is given by $s_2^n = y^n + h^n x^n + \varepsilon_2^n$, where $h^n \geq 0$ is the amount of help the parents provide for task n and x^n is the parents' ability to help the child perform activity n. As before, $x^n \sim^{iid} N\left(\mu_x^n, \sigma_x^2\right)$, is independent of all other random variables and $\mu_x^n > 0$ for all n, while $\varepsilon_2^n \sim^{iid} N(0,1)$ and is independent of all other random variables.

Finally, the parents' utility function is given by

$$u\left(\left\{s_2^n\right\}_{n=1}^N\right) = \sum_{n=1}^N g\left(s_2^n\right)$$

where $g'\left(s_2^n\right) > 0$ for all $s_2^n \in \mathbb{R}$. The parents' problem is to choose $\left\{h^n\right\}_{n=1}^N$ such that

$$\sum_{n=1}^N h^n \leq \overline{h}$$

where $\overline{h} > 0$, in order to maximize their expected utility.

Finally, let $\Psi^n\left(s_1^1, ..., s_1^N\right)$ denote the optimal help level allocated to task n, given the set of scores from the first period.

Proposition 19 *For all n, if $g''(z) > 0$ for all $z \in \mathbb{R}$, then $\Psi^n\left(s_1^1, ..., s_1^N\right)$ is increasing with s_1^n for all $\left\{s_1^{\widehat{n}}\right\}_{\widehat{n} \neq n}$.*

Proof. See appendix. ■

The idea behind the argument is the following. Any of these N different skills could be the basis of a career for the child. For example, the variable y^1 could refer to the child's musical ability, variable y^2 could refer to athletic ability and so on. By obtaining a large score in the second period on a given task n, this might increase the chances of a successful career using skill n. So, for example, s_2^1 can be the score of an entry audition for the Juilliard music school, or s_2^2 can be the performance of the child in his high school football competition. Naturally, parents would care about these signals because they are positively correlated with the child's future success.

The assumption that g is convex simply means that parents prefer the child to be very

good at one activity than average at all activities. The parents' motivation is that, if one thinks of these activities as precursors of future careers, a successful career depends on having only one very good skill (if the child becomes a professional musician, his athletic ability is pretty irrelevant in terms of his career). The consequence of this assumption is that the more convinced the parents are that a child's skill y^n is large, the more they will be willing to help the child in performing activity n.[4]

2.4.3. Matching

In most circumstances, most of the influence exerted on an agent comes from his coworkers. However, the type of coworkers an agent has is likely to depend on the agent's perceived ability. In particular, the more skilled an agent is perceived to be, the more cooperation opportunities he is likely to attract.

Consider a similar scenario as before, where there is a community of I agents all endowed with some ability level y^i, where $y^i \sim^{iid} N\left(0, \sigma_y^2\right)$. In the first period, each agent i observes public signal $s_1^i = y^i + \varepsilon_1^i$, where $\varepsilon_1^i \sim^{iid} N(0,1)$ and is independent of all other random variables. In the second period, agents randomly form pairs and must decide whether to cooperate. Say a pair (i', i'') is formed. If the agents cooperate, they receive a joint single signal $s_2 = y^{i'} + y^{i''} + \widetilde{\varepsilon}_2$, where $\widetilde{\varepsilon}_2 \sim N(0,1)$ and is independent of other random variables. If they do not, each of them receives his own signal $s_2^i = y^i + \varepsilon_2^i$ for $i = i', i''$, where $\varepsilon_2^i \sim^{iid} N(0,1)$ and is independent of other random variables.

I model the second period game between the two matched agents as follows. One of them is chosen with 50% probability to be given the opportunity to ask the other agent whether he wants to cooperate. If he has chosen to ask for cooperation, the other agent must choose whether or not to accept. Cooperation only occurs when the first agent chooses to ask for it, and the second agent accepts. Finally, I assume that each agent i's utility is equal to s_2 if there is cooperation and equal to s_2^i otherwise.

[4]Even though propositions 21 and 22 do not follow directly, given that Ψ was assumed to be a function of only one variable, it is easy to show the same exact results hold. Basically, what follows directly from these is that there is overconfidence at activity n (according to the respective definitions of propositions 21 and 22) conditional on the scores of all other activities $\left\{s_1^{\widehat{n}}\right\}_{\widehat{n} \neq n}$. By integrating out these other scores one obtains the unconditional overconfidence results of propositions 21 and 22 applied to each activity n.

Notice that, after period 1, the public belief about any agent i's ability is that

$$y^i|s_1^i \sim N\left(\frac{\sigma_y^2}{1+\sigma_y^2}s_1^i, \frac{\sigma_y^2}{1+\sigma_y^2}\right),$$

which then implies that each agent is only willing to cooperate with a partner for whom the first period's signal was positive.[5]

Proposition 20 *For any match* (i', i'')*,*

i) If $s_1^i > 0$ *for* $i = i', i''$*, the unique subgame perfect equilibrium outcome is to have cooperation.*

ii) If $s_1^i < 0$ *for either* $i = i'$ *or* $i = i''$*, the unique subgame perfect equilibrium outcome is not to have cooperation.*

Proof. Notice that each agent prefers to cooperate if he believes the other agent's expected ability is positive and prefers to refuse to cooperate if he believes it is negative. Hence, if the conditions of i) hold, the respondent strictly prefers to accept the cooperation offer and the proposer strictly prefers to propose cooperation. If the conditions of ii) hold and if a) the proposer's expected ability is negative, then the respondent will refuse cooperation, while if b) the respondent's expected ability is negative but the proposer's is positive, then the proposer refuses to propose cooperation. ■

Agents only wish to cooperate if their match is skilled because only then does the match bring value to the project. Therefore, agents who receive a high initial score will have more cooperation opportunities than those who receive a low score. In particular, from the perspective of some agent i', it is as if the second period score is equal to

$$y^{i'} + hx + \varepsilon_2$$

for some independent $\varepsilon_2 \sim N(0,1)$, where $x \sim N\left(\frac{\sigma^2}{\sigma^2+1}s_1^{i''}, \frac{\sigma^2}{\sigma^2+1}\right)$ and denotes the ability of the agent's match i'' and where h is increasing with $s_1^{i''}$ (and not a constant in the case

[5]Notice that the same would be true even if the agents had an uninformative prior. In that case, the conditional expectation of y^i would be exactly equal to s_1^i.

of $s_1^{i''} > 0$). As a result, the overconfidence results follow from the previous section.

2.5. Conclusion

In this paper, I have presented a theory of induced overconfidence. I have argued that, in many circumstances, an individual's opportunities to cooperate and receive help are positively correlated with past signals of achievement. I gave the example of an academic adviser who prefers to help students she perceives to be more skilled, parents who prefer to help their children with the skills where they already distinguish themselves and coworkers who prefer to cooperate only if they believe their partner is skilled. Overconfidence will then arise simply because those who are overconfident attract cooperation and, as a result, receive less informative signals relative to their ability when compared with those who are underconfident. The asymmetry in the cooperation opportunities is what causes the overconfidence bias in the agents' beliefs.

CHAPTER 3 : Should the Government provide public goods if it cannot commit?

3.1. Introduction

The question of who should provide public goods is an old question in the economic literature. In a seminal work, Samuelson (1954) argues that, in a world with complete information, the market will typically be unable to provide an efficient allocation of public goods due to what I refer to as the "classical" free riding problem - each agent disregards the positive impact that his private contribution to the provision of public goods has on other agents. Hence, if a benevolent dictator (BD) exists, he should be able to solve all inefficiencies simply by imposing socially optimal contributions on the agents. This line of reasoning seems to point to the superiority of the government provision of public goods.

One of the assumptions made in this argument is that this BD has complete information about the agents' preferences. Some authors, most notably Hayek (1945), see in this an argument in favor of the market. In a free market, agents make decisions based on prices, and prices contain information. Therefore, the market outcome will be more efficient as it will be a function of the agents' private information while, according to Hayek, a centralized alternative will not.

The development of the literature of mechanism design applied to the provision of public goods has analyzed the general problem of constructing mechanisms that elicit reports from the agents in order to retrieve their private information. The famous revelation principle (see, e.g. Myerson (1979)) states that one can restrict attention to revelation mechanisms, where agents simply report their private type to a mediator, which then maps those reports into allocations. By thinking of the mediator as a BD we see that the revelation principle is the answer to Hayek's argument. Any mechanism outcome (including a market mechanism) can be replicated by a BD using a revelation mechanism. However, this argument assumes that the BD is able to commit to a particular allocation of public goods even if such allocation is not optimal, given the agents' reports. In this paper, I revisit the question of who should provide public goods but assume the BD is no longer able to commit. I

specifically model the BD and assume he has preferences over the agents' utilities. In particular, I assume the BD's utility function is $W(u)$ where $u = (u_1, ..., u_N)$ is the vector of individual utilities each agent $n = 1, ..., N$ has. I also assume W is strictly increasing with each u_n, symmetric (so that all agents are cared for equally) and strictly concave. The strict concavity makes the BD prefer to impose higher transfers on the agents that value the public good the most.[1]

In the typical mechanism design approach to the problem of public good provision, the mediator (which we interpret as the BD) maps the agents' truthful reports to units of the public good to provide and to transfers each individual must make. If he has commitment power, this mapping is chosen to maximize the BD's ex-ante expected utility $E(W)$. Typically, however, such mapping will not be ex-post optimal for the BD, i.e. after he receives the truthful reports from the agents he would prefer to provide a different level of public goods and/or select different transfers to impose on each the agent. The optimal mapping chosen by a BD that is able to commit has two features that are important for this discussion: for some truthful reports, the BD will prefer to i) provide a higher quantity of the public good and, ii) alter the transfers each agent is making in order to guarantee that the agents who have a stronger desire for the public good are the ones that pay for it the most. Hence, if the BD is unable to commit to that mapping, agents will be reluctant to reveal their private information out of fear that the public good will not be provided if they do, or that they will be charged too high of a transfer.

My analysis is predicated on the assumption that the BD cannot commit. When a particular agent or institution has commitment power, it is usually understood that that agent or institution is able to write a contract. Thus, it is assumed that some other exogenous entity enforces the contract. That means that, if one of the parties breaks the contract, that outside entity will impose a harsh punishment which would make such breach undesirable - that entity is the source of authority. Hence, if a BD or a government are defined to have this authority, the contracts they write cannot be enforced by some other institution. It is

[1] Some of the results I present also hold if W is linear. I will point out when they do not in the text.

possible, however, that the BD acts as if he can commit. For example, we see that most countries have written constitutions in what appears to be an attempt from governments to gain such commitment power. However, all constitutions are changeable and many are altered quite frequently. We also constantly see promises made by governments that are quickly broken. For these reasons, it seems relatively clear to me that, at the very least, there are some limitations to the commitment power of the government.

The goal of this paper is to compare two alternative ways of providing the public good. The first one is through a BD who is unable to commit, while the second one is through the voluntary provision of public goods where the BD is absent. Given the absence of the BD in the second alternative, I label it as "Anarchy". I model these two alternatives in a similar way. I consider a model where there are two periods. In the first period, the communication period, each agent simultaneously sends a public message out of an arbitrarily large message set. The second period is where decisions about contributions are made. In the BD system, the BD imposes the contributions each agent makes as a function of the reported messages. In the anarchic system, individual agents simultaneously and voluntarily select their own private contribution as a function of their own private information as well as of the reported messages.

I make two main assumptions that are more or less standard in the public good literature: i) agents' utility functions are quasilinear, and ii) the BD must guarantee that all agents have an ex-post utility that is (weakly) higher than if the public good was not provided at all, i.e. I assume the BD faces ex-post individual rationality constraints.[2]

There are two main results from my analysis. The first result is that if the public good is binary - $g \in \{0, 1\}$ where g stands for the number of units of the public provided - all equilibrium outcomes of the BD system are also equilibrium outcomes of the anarchic system. Moreover, the opposite is not true and the expected welfare associated with some equilibria of the anarchic system is strictly higher than the expected welfare of any equilibrium of the BD system. The second result is that if $g \in \left\{0, \frac{1}{k}, ..., \frac{k}{k}\right\}$ for some integer k, there is some

[2]In the text, I justify this assumption and discuss its implications for the main results.

integer $\overline{k}$ such that the anarchic system is preferred if $k < \overline{k}$ but the BD system is preferred if $k > \overline{k}$.

The intuition is as follows. Consider the case where $g \in \{0, 1\}$ and there is complete information. In this case, there are always Bayes-Nash equilibrium outcomes of the voluntary contribution game that maximize social welfare. This is because, given that W is strictly concave, the socially optimal transfer vector is such that all agents have the same utility. Therefore, all agents' utility will be positive. That same transfer vector is also a Bayes-Nash equilibrium outcome as no agent wishes to deviate to a different transfer: by assumption, if the good is not being provided, each agent is not willing to provide it by himself; if the good is being provided, there is no incentive in making a higher transfer (for it would not increase the amount of units provided) or a lower transfer (as the good would not be provided at all which would make the agent indifferent at best). This means that the classical free-riding problem identified in Samuelson (1954) is not present when $g \in \{0, 1\}$. It is the fact that the information is private that prevents socially optimal outcomes. The first result can then be interpreted as corroborating Hayek's argument, whenever the public good provision problem is a merely informational one. However, as k increases, the classical free riding problem starts to emerge. In particular, given the quasilinear assumption on the agents' utility function, the highest level of public good that can be provided in Anarchy is $\frac{1}{k}$, as agents always have an incentive to undercut their contribution otherwise, which is the basis of the second result.

By combining the two main results, the conclusion is that there is a trade-off associated with the centralized provision of public goods. A BD is better equipped to deal with the classical free riding problem described in Samuelson (1954), but is less capable of accommodating the private information held by the agents as described in Hayek (1945). The relative strength of each of these two forces, which we measure by k, determines whether a centralized provision of public goods improves upon an anarchic one.

I believe this paper makes contributions to three different areas of the economic litera-

ture. First, to the literature on public goods. The classic literature on public good provision with incomplete information, which includes Groves (1973), d'Aspremont and Gerard-Varet (1979), Laffont and Maskin (1979) among others, typically assumes the mediator/BD has complete contracting ability. This paper relates more closely to the literature that reduces the commitment power of the BD. Schmidt (1996) provides an argument for the privatization of public firms. The idea is that, if the government is directly responsible for the firm and is unable to commit, it will receive private information that will make it less able to provide incentives for the agents employed by the firm to exert effort. Hence, the author argues, privatization (and subsequent regulation) can be seen as a useful commitment device by the government. The main difference from Schmidt (1996) to this paper is that the former focuses on the moral hazard problem rather than on the adverse selection problem the government faces.

Second, this paper may be interpreted in light of the literature on the decentralization of the government. It is possible to interpret the agents in our model as local representatives of different regions and ask the question: should the decision about a public good that affects all regions be made by a centralized government? Or should it be left to the local representatives to reach an agreement? The classical analysis of this problem is due to Oates (1972), where the author argues that decentralization will be preferred as long as the provision of the public good in a given region does not generate large enough positive spillovers on the other regions. Besley and Coate (2003) and Lockwood (2002) relax the assumption made in Oates (1972) that a centralized government selects a uniform policy for all its regions but still assume complete information. There are also several papers that analyze the same question under incomplete information but do not allow for communication among the regions (for example Kessler(2014) or Cho (2013)) which limits the benefits of decentralization. Klibanoff and Poitevin (2013) is an exception in that the authors do allow for some bargaining to occur between the regions. However, when modelling the decentralized system, it is assumed that the regions are able to celebrate contracts among themselves. As discussed above, if it is possible for the regions to celebrate contracts, then

it seems reasonable to also allow the government to celebrate contracts with the regions, which, by the revelation principle would be preferred. For this reason, in my analysis, the regions (agents) are not allowed to celebrate contracts.

Finally, our analysis of the anarchic system builds on the notion that allowing agents to communicate enhances considerably the set of allocations that form an equilibrium. Matthews and Postlewaite (1989) show that, in a bilateral trade setting, the introduction of a cheap talk stage, prior to having the traders participate in a double action, allows the implementation of a much larger set of allocations. This paper also builds on Agastya et al (2007) in that the construction of the anarchic system is very similar. The authors show that, in a public good environment with private types, quasilinear individual utility functions and a binary public good, even simple pre play communication (in that the message set is restricted) enlarges the set of allocations that are implementable through a voluntary provision game. They also show that, if we do not restrict the message set and under some conditions, the anarchic system is able to implement the expected welfare maximizing allocation, among all of those that are incentive compatible and interim individually rational. In Proposition 32 of this paper I show that the same result holds in this framework, where the welfare function is strictly concave (instead of it being linear), ex-post (rather than interim) individually rational allocations are considered and there are N agents (instead of only 2).[3] The main differences from this paper, however, are that the comparison with the BD system is absent and that we provide an analysis for the case when the public good is not binary and show how this assumption affects the success of the anarchic system as compared to the BD alternative.

The rest of the paper is organized as follows. In section 2, I formalize the setup of the model. In section 3, I describe the BD system. I characterize the welfare maximizing mechanism that could be implemented by a BD that is able to commit and analyze the consequences of eliminating that commitment power. Section 4 introduces the anarchic system and compares it with the BD system (with and without commitment power). Section

[3]In the text, I discuss in more detail why this result does not follow directly from Agastya et al (2007).

5 extends the analysis to discrete but non-binary public goods. In section 6, I discuss two extensions: first, I analyze whether the existence of a mediator could eliminate the consequences associated with the loss of commitment power by the BD; and second, I discuss the individual rationality assumption and its implications. Section 7 concludes.

3.2. Model

I consider a community with $N > 1$ agents. Each agent n is endowed with a private type $v_n \in \{\underline{v}, \overline{v}\}$ where $\overline{v} > \underline{v} > 0$. We often refer to $\overline{v}$ as the high type, and $\underline{v}$ as the low type. We assume v_n is independent and identically distributed across n and denote by $\pi \in (0,1)$ the probability that $v_n = \overline{v}$ for all n.

I assume that the public good g is either produced or not and so $g \in \{0,1\}$. The cost of providing the public good is given by $\widehat{c} \equiv Nc$, i.e. $g = \begin{cases} 1 \text{ if } \sum_{n=1}^{N} t_n \geq \widehat{c} \\ 0 \text{ otherwise} \end{cases}$, where $c \in (\underline{v}, \overline{v})$ denotes the average cost of providing the public good. I assume $c \in (\underline{v}, \overline{v})$ so that, at least for $v = (\overline{v}, ..., \overline{v})$ it is at least possible to finance the public good, and for $v = (\underline{v}, ..., \underline{v})$ it is not. The utility function of each agent n is given by $u_n = v_n g - t_n$ - it depends on whether the public good is provided and on the transfer $t_n \in \mathbb{R}$ agent n makes.

Allocations are evaluated based on a welfare function $W : \mathbb{R}^N \to \mathbb{R}$ - a function of the individual utilities $u = (u_1, ..., u_n)$. I normalize $W(0, .., 0) = 0$ and assume that W is twice continuously differentiable, strictly increasing, strictly concave and symmetric.

I also assume $\overline{v} < \underline{v} + c$. This assumption is convenient for two reasons. First, it implies that $\overline{v} < \widehat{c}$ which means that an individual is unable to provide the public good by himself. Second, as I argue in the next section, the assumption guarantees that, in the first best allocation to be defined, there are no negative transfers.

3.3. Benevolent Dictator

In this section, I assume the existence of a benevolent dictator (BD) who is responsible for the provision of the public good. I assume the BD's interests are aligned with those of the society and so, the BD's utility function is given by W.

There are two periods in this framework. In the first period, the communication period, each agent n simultaneously sends a message m_n from an arbitrarily large message set M_n. At the end of the first period, all messages become known to all agents, including the BD. In the second period, the BD selects whether the good is provided or not - $g \in \{0,1\}$ - and a transfer scheme $t = (t_1, ..., t_N)$, should the good be provided, such that $\sum_{n=1}^{N} t_n \geq \widehat{c}$.[4] I assume that the BD may not inflict a loss on any agent as a result of financing of the public good, i.e. I impose the BD faces ex-post individual rationality constraints.[5]

3.3.1. With Commitment Power

I start with the more traditional analysis of the BD's problem when he has commitment power - when he can commit not to alter the ex-ante optimal maps from the messages sent by the agents to his own actions. Given the BD is able to commit, we can refer to the Revelation Principle and restrict our attention to revelation mechanisms - $M_n = \{\underline{v}, \overline{v}\}$ for all n - where truthful reporting is an equilibrium. The BD's problem is then to choose $\rho : \{\underline{v}, \overline{v}\}^N \to [0,1]$ - the probability that the public good is provided, and $t : \{\underline{v}, \overline{v}\}^N \to \mathbb{R}^N$ - the contributions he demands from each agent should the good be provided, given any possible truthful reports.[6] Notice that lotteries over transfers are not optimal given that W is strictly concave with u_n while u_n is linear with t_n. Hence, the objective function of the BD is

$$V(\rho, t) = \sum_{v \in \{\underline{v}, \overline{v}\}^N} \Pr\{v\} \rho(v) W(v_1 - t_1(v), ..., v_N - t_N(v)) \tag{3.1}$$

[4] If the public good is not provided I set $t = (0, ..., 0)$.

[5] I discuss the implications of this assumption in section 6.2.

[6] This formulation assumes that it is only possible to impose transfers on the agents when the good is provided.

The BD also faces incentive constraints,

$$\sum_{v_{-n}\in\{\underline{v},\overline{v}\}^{N-1}} \Pr\{v_{-n}\}\,\rho(v_n,v_{-n})\,(v_n - t_n(v_n,v_{-n})) \tag{3.2}$$
$$\geq \sum_{v_{-n}\in\{\underline{v},\overline{v}\}^{N-1}} \Pr\{v_{-n}\}\,\rho\left(v'_n,v_{-n}\right)\left(v_n - t_n\left(v'_n,v_{-n}\right)\right)$$

for all v'_n, v_n and n

ex-post participation constraints,

$$\rho(v)\,(v_n - t_n(v)) \geq 0 \text{ for all } n \text{ and } v \tag{3.3}$$

and feasibility constraints

$$\rho(v)\left(\sum_{n=1}^{N} t_n(v) - \widehat{c}\right) \geq 0 \text{ for all } v \tag{3.4}$$

Condition (3.2) imposes that all agents prefer to truthfully report, rather than misreport their type. Condition (3.3) guarantees that all agents have a positive ex-post utility, no matter what the realization of v is. This condition effectively grants veto power on each agent as no allocation that makes an agent worse off is allowed. Condition (3.4) guarantees that the public good is fully funded by the agents whenever it gets provided.

We start by characterizing the first best allocation (ρ^*, t^*) - where we maximize (3.1) subject to (3.4). Notice that it is efficient to provide the public good if and only if the sum of the valuations of all agents exceeds the cost of providing it. Given the type space of each agent is binary, it is efficient to provide the public good if and only if there are enough high valuation agents. Notice also that, if v is such that $\rho^*(v) = 0$, the choice of the transfer vector is irrelevant so, WLOG, we set $t_n^*(v) = c$ for all n.

Let $i(v)$ be the number of high reports in v ($i(v) \equiv \sum_{n=1}^{N} 1\{v_n = \overline{v}\}$) and $\widehat{i} \in \mathbb{N}$ be the smallest number of high type agents for which it is efficient to provide the public good (for

all integers $i \geq \widehat{i}$, $i\overline{v}+(N-i)\,\underline{v} \geq \widehat{c}$). For simplicity, we assume $\widehat{i}\overline{v}+\left(N-\widehat{i}\right)\underline{v} > \widehat{c}$ - if there was complete information the BD would never be indifferent - which allows for a simpler exposition of the results.

Proposition 21 *For all v and n,*

$$i)\ \rho^{*}(v) = \begin{cases} 1 \text{ if } i(v) \geq \widehat{i} \\ 0 \text{ if } i(v) < \widehat{i} \end{cases} \text{ and}$$

$$ii)\ t_{n}^{*}(v) = \begin{cases} \overline{v} - \frac{i(v)\overline{v}+(N-i(v))\underline{v}}{N} + c \text{ if } v_n = \overline{v} \text{ and } \rho^{*}(v) > 0 \\ \underline{v} - \frac{i(v)\overline{v}+(N-i(v))\underline{v}}{N} + c \text{ if } v_n = \underline{v} \text{ and } \rho^{*}(v) > 0 \end{cases}$$

Proof. If $\rho^{*}(v) > 0$ then

$$t^{*}(v) \in \arg \max_{t \in \mathbb{R}^{N}: \sum_{n=1}^{N} t_n \geq \widehat{c}} W(v_1 - t_1, ..., v_N - t_N)$$

which implies $ii)$ given the strict concavity of W. Notice also

$$\max_{t \in \mathbb{R}^{N}: \sum_{n=1}^{N} t_n \geq \widehat{c}} W(v_1 - t_1, ..., v_N - t_N) \geq W(0, ..., 0)$$

if and only if $i(v) \geq \widehat{i}$. ■

In the first best allocation, the public good always gets provided as long as it is efficient to do so. This is because if it is efficient to provide the public good, it is always possible to find transfers that make all agents better than what they would have been had the public good not been provided.

Given that W is strictly concave, the BD has equality concerns. This implies that, conditional on each v, the BD selects transfers in order to equate all agents' utilities. Given that the agents have transferable utility, transferring utility from an agent with a high utility to an agent with a low utility, through a reallocation of the transfers, is welfare increasing. Hence, $t^{*}(v)$ is such that all agents have similar utilities given v. The assumption that

$\overline{v} < \underline{v} + c$ is used in this proposition as it guarantees that the BD does not choose a negative transfer as a way to redistribute utility.[7]

Now, I analyze the original problem the BD faces: to maximize (3.1) subject to (3.2), (3.3) and (3.4). Even though there may be multiple solutions, I only discuss solution $\left(\overline{\rho}, \overline{t}\right)$, that is characterized by the fact that it is the only solution where the probability of providing the public good $\overline{\rho}$ only depends on the number of high and low agents, and not on their identity.[8] Formally, $\overline{\rho}$ is such that $\overline{\rho}(v) = \overline{\rho}(v')$ if $i(v) = i(v')$ for all v, v'.

The problem of finding $\left(\overline{\rho}, \overline{t}\right)$ is a fairly standard one. The only novelty has to do with the fact that W is strictly concave rather than linear, which makes it harder to implement efficient allocations as I now explain. If the BD was to impose the first best solution (ρ^*, t^*), low valuation agents would report truthfully, given that the transfer they would have to pay, if they misreported, would be higher than their valuation. However, high type agents might not. Hence, in order to provide incentives for high type agents to report truthfully, the BD must either decrease the transfer that high type agents make and/or decrease the probability the public good is provided if too many agents report low. If the BD is inequality averse, there is a welfare cost associated with the former. Hence, the BD will prefer to distort ρ more than what it would have if W was linear.

Proposition 27 characterizes $\left(\overline{\rho}, \overline{t}\right)$.

Proposition 22 *i)* $\sum_{n=1}^{N} \overline{t}_n(v) = \widehat{c}$,

ii) $\overline{t}_n(v) = \overline{t}_{n'}(v)$ *if* $v_n = v_{n'}$ *for all* $v_n, v_{n'}, v$.

iii) $\overline{t}_n(v_n, v_{-n}) = \overline{t}_n\left(v_n, v'_{-n}\right)$ *if* $i(v_n, v_{-n}) = i\left(v_n, v'_{-n}\right)$ *for all* v_n, v_{-n}, v'_{-n}.

iv) $\overline{t}_n(v) \leq t_n^*(v)$ *if* $v_n = \overline{v}$ *and* $\overline{t}_n(v) \geq t_n^*(v)$ *if* $v_n = \underline{v}$ *with both inequalities being strict if* (ρ^*, t^*) *is not incentive compatible.*

[7] If $\overline{v} < \underline{v} + c$ then $\underline{v} + c > \frac{i(v)\overline{v} + (N - i(v))\underline{v}}{N}$ for all v and N.

[8] The multiplicity of solutions has two reasons. First, if $\rho(v) = 0$ the decision on $t(v)$ is irrelevant. Following the same convention as with (ρ^*, t^*), WLOG, I set $\overline{t}_n(v) = c$ for all n whenever $\overline{\rho}(v) = 0$. Second, there may be degrees of freedom in the decision of $\rho(v)$ for the set of v that have a common $i(v)$. By imposing that $\rho(v) = \rho(v')$ if $i(v) = i(v')$ for all v, v' we obtain $\overline{\rho}$.

v) There is $\bar{i} \in \left\{\widehat{i}, ..., N-1\right\}$ *such that,* $\overline{\rho}(v) = \begin{cases} 1 \text{ if } i(v) > \bar{i} \\ 0 \text{ if } i(v) < \bar{i} \end{cases}$.

Proof. See appendix. ■

As I mention in the Introduction, and show in section 4, it is possible to implement $(\overline{\rho}, \overline{t})$ in Anarchy. Property i), that states that there are no wasted transfers, plays a key role in that argument. I will discuss that role in more detail in section 4. Properties ii) and iii) taken together imply that transfers are anonymous, i.e. an agent's transfer only depends on his type and on how many high type agents there are; not on his index n. This is guaranteed by the strict concavity of W and is a property of any solution of the problem. Property iv) states that there is more inequality among the agents than in the first best. As discussed above, the BD is forced to distort the optimal transfers it would have wanted to implement in order to make high valuation agents less willing to misreport. Given that the distortion decreases the transfers imposed on high valuation agents, this means that low valuation agents are now contributing more and so the inequality is higher. Finally, property v) states that there is less provision of the public good than the efficient one for similar reasons. In order to create incentives for high valuation agents to report truthfully, if there are too many agents reporting to have low valuations, the public good is not provided, even though it would have been efficient to.

The last two properties also imply that the BD solution may not be ex-post welfare maximizing. After knowing the agents' truthful reports, the BD would rather provide the public good if and only if it was efficient to do so and impose the ex-post optimal transfers. But knowing this, agents would be reluctant to truthfully report in the first place. In the next section, I describe the consequences of removing the commitment power of the BD.

3.3.2. Without Commitment Power

In this section, I study what would happen if the BD did not have commitment power. The framework is the same as before except that, after the first period and once all messages have been revealed, the BD has no choice but to behave optimally, given the beliefs he has

at that time. As such, it is no longer possible to refer to the revelation principle, because the beliefs the BD holds at the beginning of the second period constrain his decision. I refer to this mechanism (and not any other mechanism that requires commitment power by the BD) as the BD system in the remainder of the paper and it is this system that I compare with the anarchic system, to be defined in the next section.

Let $\sigma_n = \left(\sigma_n^{\overline{v}}, \sigma_n^{\underline{v}}\right)$ denote a generic strategy of agent n where $\sigma_n^{v_n} \in \nabla\left(M_n\right)$ for all $v_n \in \{\underline{v}, \overline{v}\}$ and let $\sigma = (\sigma_1, .., \sigma_N)$. An agent's strategy is a choice of a probability distribution over the message space for each type.

Let $\xi = \left\{\xi^m : m \in M \equiv \{M_n\}_{n=1}^N\right\}$ denote a generic strategy of the BD where $\xi^m \in \nabla\left(\mathbb{R}^N\right)$. A BD's strategy is a choice of a probability distribution over the set of all possible transfer vectors for each message vector received. Notice that I do not include a choice over the probability that the public good is provided. In principle, it could be possible for the BD to randomize between providing and not providing the public good, given some message vector m, which would require the BD to specify a probability the public good is provided and a transfer vector in case it did, just like in the previous section. However, if the BD cannot commit, he must be indifferent between providing and not providing the public good to be able to do this randomization. As I argue below, this does not happen and the BD always strictly prefers one of the two options. Therefore, WLOG, it is enough to specify a distribution over the transfer vector with the understanding that these are no longer contingent transfers, which means that the public good is provided if and only if the sum of the transfers exceeds the cost of providing the public good.

A BD equilibrium is a strategy profile (σ, ξ) and a set of beliefs that form a perfect bayesian equilibrium (PBE): i) given the beliefs that follow message m, the choice of ξ^m is optimal; ii) for all n, σ_n is chosen optimally anticipating all agents' future play, iii) beliefs are updated according to Bayes' rule whenever possible.

If $\left(\overline{\rho}, \overline{t}\right) = (\rho^*, t^*)$ (if the first best allocation is incentive compatible), then a BD without commitment power can also implement $\left(\overline{\rho}, \overline{t}\right)$. This is because, if agents report truthfully,

it will be in the BD's best interest, given his beliefs at the end of the first period, to act as specified by $\left(\overline{\rho},\overline{t}\right)$. However, if $\left(\overline{\rho},\overline{t}\right) \neq (\rho^*, t^*)$, things are not as simple. Notice that, in this case, truthful reporting is no longer an equilibrium for otherwise (ρ^*, t^*) would be incentive compatible. Hence, there is no equilibrium that is fully informative. The ability to obtain information from the agents is therefore hindered by the lack of commitment power. In proposition 28, I show that this difficulty in obtaining information from the agents generates a strict loss in the expected welfare of society, associated with the loss of commitment power.

Proposition 23 *If* $\left(\overline{\rho},\overline{t}\right) \neq (\rho^*, t^*)$ *the value of the BD system (the highest expected welfare among all BD equilibria) is strictly smaller than* $V\left(\overline{\rho},\overline{t}\right)$.

Proof. Recall that, for any solution of the problem of the previous section (where the BD had commitment power), the only incentive constraint that binds is one that imposes that the sum across all agents of the expected utility of reporting truthfully, when they have a high valuation, is higher than misreporting - see the proof of Proposition 27 for details. Therefore, it follows that, for any solution of the problem of the previous section, we have that, when all agents' types are high, the public good is provided with probability 1 and each agent makes a transfer of c, just like with $\left(\overline{\rho},\overline{t}\right)$.

Suppose such solution is implementable by a BD without commitment power and take any BD equilibrium that implements it. Consider any message $m = (m_1, ..., m_N)$ such that, for all n, m_n is played by type $\overline{v}$ of agent n. If that solution is implementable, it must be that after m, the BD provides the public good and the transfer agent n must make is equal to c. However, that is only possible if, for all n, m_n is sent **only** by type $\overline{v}$ of agent n, for otherwise agent n's maximum transfer would be $\underline{v} < c$ (given that the BD is unable to impose a transfer on any agent that makes his ex-post utility negative). This implies that there must be truthful reporting by each agent. But that cannot be given that $\left(\overline{\rho},\overline{t}\right) \neq (\rho^*, t^*)$. ■

I now provide a characterization of the set of BD equilibria. Notice that, for any

BD equilibrium (σ, ξ), ξ reflects the optimal decision the BD makes, at the beginning of the second period, and given his beliefs about the agents' types. Such beliefs depend on σ - the reporting strategies of the agents - and on m - the realized vector message. Let $p^{\sigma}(m) \in [0,1]^N$ be such that $p_n^{\sigma}(m)$ represents the probability that the BD places on agent n being of type $\overline{v}$ after observing message m and given reporting strategy σ. It is also convenient to define $i^{\sigma}(m)$ to be the number of agents the BD is certain are of type $\overline{v}$ ($i^{\sigma}(m) \equiv \sum_{n=1}^{N} 1\{p_n^{\sigma}(m) = 1\}$) and $x^{\sigma}(m)$ to be the optimal transfer scheme for the BD, after observing message m and conditional on providing the public good, i.e. $x^{\sigma}(m) = \arg\max_{x \in \mathbb{R}_+^N} E_{v|\sigma,m}\left[W(v_1 - x_1, ..., v_N - x_N)\right]$ subject to the ex-post individual rationality constraints ($x_n \leq \underline{v}$ if $p_n^{\sigma}(m) < 1$) and feasibility constraints ($\sum_{n=1}^{N} x_n \geq \widehat{c}$).

Lemma 3 *For all BD equilibrium (σ, ξ) and for all $m \in M$,*

i) If $i^{\sigma}(m) \geq \widehat{i}$ then $\xi^m(a) = \begin{cases} 1 \text{ if } a = x^{\sigma}(m) \\ 0 \text{ otherwise} \end{cases}$,

ii) If $i^{\sigma}(m) < \widehat{i}$ then $\xi^m(a) = \begin{cases} 1 \text{ if } a = (0, ..., 0) \\ 0 \text{ otherwise} \end{cases}$,

iii) If $p_{n'}^{\sigma}(m') > p_{n'}^{\sigma}(m'')$ and $p_n^{\sigma}(m') = p_n^{\sigma}(m'')$ for all $n \neq n'$, then $x_{n'}^{\sigma}(m') \geq x_{n'}^{\sigma}(m'')$. The last inequality is strict if $x_{n'}^{\sigma}(m') < \underline{v}$,

iv) $x_n^{\sigma}(m) < \underline{v}$ if $p_n^{\sigma}(m) = 0$.

Proof. Properties i) and ii) follow from the fact that, if $p_n^{\sigma}(m) < 1$, then $x_n^{\sigma}(m) \leq \underline{v}$ given the ex-post individual rationality constraint the BD faces. Hence, the maximum revenue gathered is given by $i^{\sigma}(m)\overline{v} + (N - i^{\sigma}(m))\underline{v}$ which implies ii). If $i^{\sigma}(m) \geq \widehat{i}$, then it is optimal for the BD to provide the good given that W is strictly increasing and that u_n is linear with t_n for all n. Property iii) follows from the strict concavity of W. The weak inequality becomes strict if the ex-post participation constraint for player n' does not bind (and holds with an inequality) for message m'. Property iv) comes from the fact that the highest transfer agent n has to pay if $p_n^{\sigma}(m) = 0$ is when $i^{\sigma}(m) = \widehat{i}$ and, for all n' such that $p_{n'}^{\sigma}(m) < 1$ it is the case that $p_{n'}^{\sigma}(m) = 0$. In that case, m fully reveals the true v, and so the transfers that follow are such that all agents have a strictly positive utility which

implies $x_n^{\sigma}(m) < \underline{v}$. ■

Notice that, given the ex-post individual rationality constraints, the BD cannot impose a transfer above $\underline{v}$ on an agent unless he is certain that the agent has a high valuation. This is because, if not, a low type agent could have a negative ex-post utility. Therefore, it is only possible to fund the good if there are at least $\widehat{i}$ "certified" high valuation agents. Given that the BD is inequality averse, he will select a transfer vector that makes high valuation agents pay for most of the good, should he choose to provide it. This creates an additional incentive for agents to misreport, as high valuation agents are reluctant to announce their type knowing the BD's intentions. Notice also that, regardless of his beliefs, the BD is never indifferent - he always selects a unique transfer vector in the second period. This uniqueness is due mainly to the strict concavity of W, which makes it such that $x^{\sigma}(m)$ is unique for all σ and m, and also to the assumption made above that $\widehat{i}\overline{v} + \left(N - \widehat{i}\right)\underline{v} > \widehat{c}$.

Now, I analyze the behavior of the agents. As discussed above, the loss of commitment power makes extracting information from the agents harder. As a result, typically we should not expect truthful reporting from an equilibrium (unless the first best is incentive compatible) and, as result, there should be some amount of pooling in any equilibrium.[9]

In this framework, a high type agent is reluctant to reveal his type because he knows that it will cause him to make a higher transfer. However, a high report also makes it more likely the public good is provided. High valuation agents, by definition, value more the provision of the public good, so they will be more willing to make a high report than low valuation agents. This means that, even though full separation between the agents cannot be achieved if there is no truthful equilibrium, some may.

Consider a strategy profile for each agent n where only two messages are played. Label those messages as H and L. Let the strategy of the low type for any agent n to be to only send message L while the high type sends message L with probability $s_n \in [0,1]$

[9]See, for example, Klibanoff and Poitevin (2013), where a similar problem is studied under the assumption that agents have continuous types. It is shown that there is no truthfull equilibrium and an equilibrium is characterized where each agent reveals only a range from where his type belongs to.

and H with probability $1-s_n$. We refer to this strategy profile for the agents as $\widetilde{\sigma}(s) = (\widetilde{\sigma}_1(s_1), ..., \widetilde{\sigma}_N(s_N))$.

Proposition 24 *There is $s \in [0,1]^N$ and $\widetilde{\xi}$ such that $\left(\widetilde{\sigma}(s), \widetilde{\xi}\right)$ is a BD equilibrium and induces the highest expected welfare among all BD equilibrium profiles.*

Proof. See appendix. ■

The BD equilibrium described above gives the choice to high valuation agents to reveal their type by sending message H or to hide it by sending message L. High valuation agents are indifferent between the two choices because, while the former leads to a higher provision of the public good, the latter involves a smaller transfer.

Finally, there is one last feature of the BD system I want to highlight. Notice that there is always some information transmission in the BD equilibrium that maximizes expected welfare, i.e. the s referred to in proposition 30 is not $(1, ..., 1)$. This is because having exactly $\widehat{i}$ agents truthfully reporting (reporting H when their valuation is high and L when their valuation is low) while all other agents send a meaningless message (report L regardless of type) is part of a BD equilibrium given that none of those $\widehat{i}$ agents wants to misreport: if an agent has a high valuation, reporting L would lead to the public good not being provided at all; while if the agent has a low valuation, reporting H would lead to a transfer that would exceed $\underline{v}$ in every event the public good gets provided. This equilibrium is preferred to the "no communication" BD equilibrium given that all agents have a strictly positive expected utility, which is clearly an improvement to not providing the public good at all. The implication of this feature is that the BD's behavior is responsive to the private information held by the agents. Hence, it is not necessarily the case that a centralized government selects a "one size that fits all" policy as sometimes is assumed in some of the decentralization literature, in particular in Oates (1972). Nevertheless, the general argument in favor of decentralization still follows in our framework because, even though in a BD system there is still private information being utilized in the BD's decision, the amount of information transmitted is inferior to that of a decentralized system as I argue

in the next section.

3.4. Anarchy

In this section, I consider what would happen if the BD was not involved in the provision of public goods - if all agents were free to decide their own voluntary contribution in Anarchy. I model this new system in a similar way as the system with a BD. There are still two periods. The first one is a communication period, where each agent simultaneously sends a message $m_n \in M_n$ and all messages become publicly available at the end of the period. The difference now is that, in the second period, agents individually and simultaneously choose their own contribution to the public good, rather than having the BD impose its decision. If the sum of the contributions exceeds $\widehat{c}$ the public good is provided; otherwise it is not.

A strategy profile for the agents is now composed of two elements - a reporting strategy profile which is still denoted by σ - and a contribution strategy profile $\phi = (\phi_1, ..., \phi_N)$ where $\phi_n = \{\phi_n^{v_n,m} : v_n \in \{\underline{v}, \overline{v}\} \text{ and } m \in M\}$ and $\phi_n^{v_n,m} \in \nabla(\mathbb{R}_+)$. An agent's contribution strategy is a choice of a probability distribution over the set of possible individual transfers, given the agents' type and reported messages.

An anarchic equilibrium is a strategy profile (σ, ϕ) and a set of beliefs that form a PBE: i) given any message and subsequent beliefs, ϕ induces a Bayes-Nash equilibrium of the second period game; ii) for all n, σ_n is chosen optimally anticipating all agents' future play, iii) beliefs are updated according to Bayes' rule whenever possible.

The first result in this section concerns the connection between the BD system and the anarchic system. At first sight, it would appear as though the BD system should have an advantage, as it does not require all agents to be willing to provide contributions for the public good - it can simply enforce those transfers (notwithstanding the veto power each agent has in the form of the ex-post individual rationality constraints). However, in the context of binary public goods, such advantage ends up not mattering, as it is in the agents' best interest to select the transfers a BD would have selected.

Consider any BD equilibrium (σ, ξ) and let $\phi(\xi)$ be such that

$$\phi_n^{\overline{v},m}(\xi)(a) = \xi_n^m(a) \text{ for all } a, m, n$$

and

$$\phi_n^{\underline{v},m}(\xi)(a) = \begin{cases} \xi_n^m(a) \text{ if } x_n^\sigma(m) \leq \underline{v} \\ 1 \text{ if } a = 0 \text{ and } x_n^\sigma(m) > \underline{v} \\ 0 \text{ otherwise} \end{cases} \text{ for all } m, n$$

Now consider the anarchic strategy profile $(\sigma, \phi(\xi))$. In this profile, agents report as they did in the BD system and then choose the exact same transfers the BD would have selected, in the path of play - recall that, in any BD equilibrium (σ, ξ), it must be that $x_n^\sigma(m) \leq \underline{v}$ whenever $p_n^\sigma(m) < 1$, which means that the only event where low type agent n does not choose the same transfer as the BD would have is when he has deviated in the first period (played a message such that $\sigma_n^{\underline{v}}(m) = 0$). Hence, this anarchic strategy profile induces the same outcome as the BD equilibrium (σ, ξ). In proposition 31, I show that $(\sigma, \phi(\xi))$ is indeed an anarchic equilibrium.

Proposition 25 *For all BD equilibrium (σ, ξ), $(\sigma, \phi(\xi))$ is an anarchic equilibrium.*

Proof. Consider any (σ, ξ) BD equilibrium and the corresponding anarchic profile $(\sigma, \phi(\xi))$.

First, notice that $\phi(\xi)$ induces a Bayes-Nash equilibrium of the voluntary contribution game that follows any message m. This is because the BD's choice reflected in ξ does not involve any randomization over the transfer vectors. For each m, the BD chooses one and only one transfer vector - see Lemma 4. This means that the only strictly beneficial deviation from an agent would be to make a 0 transfer. Given ξ, only low types would have such incentives, which is accommodated in the definition of $\phi(\xi)$.

The last property $(\sigma, \phi(\xi))$ has to have is that no agent wants to misreport. Because (σ, ξ) is a BD equilibrium and because, on the path of play, no agents deviates from the transfer the BD would have chosen, we know that it can only be beneficial for an agent to misreport, if that agent also deviates in the second period. High valuation agents always

select the transfers specified in ξ so they will not want to misreport. However, low valuation agents, after some messages, will prefer to select a transfer of 0 rather than the transfer specified by ξ. To complete the proof, I show that low valuation agents do not misreport.

Suppose low valuation agent n misreports and sends any message m_n such that $\sigma_n^v(m_n) = 0$. If m_n is such that $p_n^\sigma(m_n, m_{-n}) < 1$ for all $m_{-n} \in M_{-n}$, then that agent's decision in the second period would be to select the transfer the BD would have selected, which implies that such deviation is not strictly preferred. The only other alternative is that $p_n^\sigma(m_n, m_{-n}) = 1$ for all $m_{-n} \in M_{-n}$. In that case, the transfer the agent would select would be 0 and the public good would not be provided. Hence, the expected utility of deviating to such message would be 0, which is weakly smaller than playing according to σ_n^v. ■

For $(\sigma, \phi(\xi))$ to be an anarchic equilibrium, it is necessary that agents do not deviate in neither period. Typically, agents will not want to deviate in the second period. Recall that, after receiving all messages, the BD decides one and only one transfer vector. Even if an agent had the opportunity to choose a different transfer for himself he would not as higher transfers would not lead to a higher provision of the good and a 0 transfer leads to a 0 utility. It follows that if agents do not want to deviate in the second period, they also will not want to deviate in the first period, for otherwise they would have also deviated in a BD system and (σ, ξ) would not be a BD equilibrium.

It is also important to point out that this result does not depend on the linearity assumption of the agents' utility functions, nor on the private types assumption. What is important is that, for given beliefs, the BD always has a strict preference for some transfer vector and prefers to make agents with high valuations the primary contributors for the public good (which, in my framework, is guaranteed by the strict concavity of W).[10]

If $(\sigma, \phi(\xi))$ is an anarchic equilibrium, then all BD equilibria can be "replicated" in Anarchy. As I show in the next section, this is a special feature of the binary public good but an important one nonetheless. First, because there are many examples of interest of

[10] If W was linear, as in Agastya et al (2007), this result would not be true.

binary public goods and the discussion of how to fund such projects is very much ongoing.[11] And second, and perhaps more importantly, because it highlights the true benefit of having government intervention in the provision of public goods. A government is valuable not because it is more able to elicit information from the agents but because of the "classical" free riding problem. But not all public goods have the "classical" free riding problem. Binary public goods do not, which is why an anarchic provision of the public is superior.

Finally, it is interesting to note that the argument sketched in the Introduction, that said that a BD can always replicate the outcome of an anarchic system, by referring to the revelation principle, is completely reversed once we remove the commitment power. If the BD is unable to commit, it is the anarchic system that is able to replicate any outcome generated by the BD system.

The next result qualifies the anarchic system in the context of incentive compatible, ex-post individually rational mechanisms.

Proposition 26 *If* $M_n = \{\overline{v}, \underline{v}\} \times [0,1]$*, there is an anarchic equilibrium* $\left(\widehat{\sigma}, \widehat{\phi}\right)$ *that implements* $\left(\overline{\rho}, \overline{t}\right)$*.*

Proof. See appendix. ■

Proposition 32 states that the optimal incentive compatible and ex-post individually rational allocation is implementable in Anarchy. The anarchic equilibrium $\left(\widehat{\sigma}, \widehat{\phi}\right)$ is described in more detail in the proof of Proposition 32, but the idea is to have agents truthfully report their type in the first period. This means that, at the beginning of the second period, the realization of v would be known to all agents. In the second period, agents then select transfers according to $\overline{t}$ whenever $\overline{\rho}(v) = 1$ and select a transfer of 0 if $\overline{\rho}(v) = 0$.

The reason why such an anarchic equilibrium is possible is similar to Proposition 31. Typically, no agent wants to deviate from the transfer specified by $\left(\overline{\rho}, \overline{t}\right)$ on the second period, and because $\left(\overline{\rho}, \overline{t}\right)$ is incentive compatible, no agent wishes to misreport. The only additional difficulty that allocation $\left(\overline{\rho}, \overline{t}\right)$ brings is that it is possible, for some v, that

[11] See, for example Barbieri and Malueg (2010), Makris (2009) among many others.

$\overline{\rho}(v) \in (0,1)$. One way to try to replicate such $\overline{\rho}(v)$ would be to make agents randomize in their transfer decisions. However, that option would waste resources given that, for some randomizations it could be that the sum of contributions exceeds $\widehat{c}$ or is between 0 and $\widehat{c}$. Instead, by extending the message space to incorporate a report over $[0,1]$ it is possible to have the agents coordinate play. By coordinating, agents only provide the public good with probability $\overline{\rho}(v)$ and, when they do, they select the proper transfers. I explain in more detail how we can go from the reports over $[0,1]$ to attain coordination among the agents in the proof of Proposition 32, but, for all purposes, it is equivalent to having a public signal distributed uniformly in $[0,1]$.[12] If the public signal is smaller than $\overline{\rho}(v)$ the agents coordinate in making positive transfers that lead to the provision of the public good, while if not they coordinate in not making any transfers.

In Agastya et al (2007) a similar result is derived, where the authors show that an anarchic system can implement the welfare maximizing allocation among all incentive compatible and interim individually rational allocations. The fact that the individual rationality constraint is imposed on the interim level and that the welfare function is linear allows the problem of finding that optimal allocation to be simplified to the search for ρ only, as transfers can always be found to make the allocation both incentive compatible and interim individually rational. The authors show that the problem of finding such ρ in this context always leads to $\rho(v) \in \{0,1\}$ for all v. The main difficulty in that framework is to find the appropriate transfers than can be voluntarily be given in Anarchy.

In my framework, however, it is not enough to restrict the search of the optimal allocation only to ρ, given that transfers enter non-linearly in the welfare function and there are more individual rationality constraints to be satisfied. Hence, the difficulty is not to find the transfers to be given in Anarchy (as they come directly from solving the BD with commitment power problem) but rather how to deal with the fact that now there may be v such that $\rho(v) \in (0,1)$, which is why agents must report on more than just their type, as

[12] The idea is that it is possible to create "jointly controlled lotteries" following Matthews and Postlewaite (1989) and, originally, Aumann et al (1968).

I describe above.

The implications of proposition 32 are simple. The anarchic system is optimal among the set of incentive compatible and ex-post individual rational mechanisms. This fact gains an even higher importance as Anarchy is costless - it simply involves agents communicating and then making their own decisions. It may be that other systems are not. Any system where a government has a role involves some costs associated with paying the government's employees. So even if, somehow, the government could commit to particular allocations, it would still generate a smaller expected welfare if such costs were accounted for.

A key assumption in the results of Propositions 31 and 32 is the binary nature of the public good. In the next section, I argue that the BD's role becomes much more important when this assumption is relaxed.

3.5. Discrete Public Good - general case

In this section, I make only one change with respect to the previous framework. I now assume that $g \in \left\{0, \frac{1}{k}, \frac{2}{k}, .., \frac{k}{k}\right\}$ for some $k \in \mathbb{N}$ - even though g is still discrete it is no longer binary. The cost of producing g units of the public good is $g\widehat{c}$ while the benefit for agent n of type v_n is gv_n.

This context brings about a new problem: a "classical" free riding problem as identified in Samuelson (1954). Now, even if there is complete information, the ex-post optimal transfer vector may not be a Nash equilibrium of the voluntary contribution game. For any v, the ex-post optimal decision is to provide the public good in 1 unit whenever $i(v) \geq \widehat{i}$ and set up some transfers that make high valuation agents make higher transfers. However, even if v was commonly known, there would be no Bayes Nash Equilibrium that would provide more than $\frac{1}{k}$ units of the public good. This is because, for more than $\frac{1}{k}$ units to be provided, someone must be making a transfer of at least $\frac{2}{k}c$. That someone would be better off by decreasing his transfer by $\frac{2}{k}c$ given that that would lead, at most, to a decrease of $\frac{1}{k}$ units of the public good. Given that $2c > \overline{v}$ the agent prefers that outcome.

This observation leads to the following result:

Proposition 27 *The value of the anarchic system (the highest expected welfare among all anarchic equilibria) is decreasing with k and converges to* 0.

Proof. In the appendix, I show that the value of the anarchic system is decreasing with k. It converges to 0 because the highest utility any given agent can obtain is bounded above by $\overline{v}\frac{1}{k}$ which converges to 0 as k increases. ■

In this context, k may be interpreted as a measure of the "classical" free riding problem. Proposition 33 states that, as that "classical" free riding problem becomes more severe, the anarchic system becomes more ineffective.[13] Proposition 34 states that, in contrast, the BD system is unaffected by how large k is.

Proposition 28 *The value of the BD system (the highest expected welfare among all BD equilibria) is independent of k.*

Proof. In the appendix, I show that the BD's decision, for any given beliefs, is always to provide either 1 or 0 units of the public good. Hence, his decision is independent of k. ■

Proposition 34 follows from the linearity assumption on both the utility functions of the agents and on the cost structure imposed. Basically, if providing $y < 1$ units of the public good is preferred to not provide it all all, providing 1 unit of the public good will be optimal, as we are just scaling up the benefits of the agents.

By combining the two previous propositions and using the fact that there is a BD equilibrium with a strictly positive expected welfare (one where exactly $\widehat{i}$ agents report truthfully while the rest sends a single message regardless of type), the following result follows.

Proposition 29 *There is some natural number $\overline{k}$ such that the value of the anarchic system is higher than the value of the BD system if and only if $k \leq \overline{k}$.*

[13] In a similar context, Barbieri (2012) also highlights some of the problems the voluntary provision of public goods with pre-play communication may have with non-binary goods.

Proposition 35 highlights the relative virtues of each of the two systems. Anarchy is more effective when dealing with "informational" free-riding, as agents are more willing to share private information when they know there is no BD who has the power to enforce transfers against their will. The BD system's virtues lie on the reduction of "classical" free-riding. The parameter k measures the relative importance of each of the two free-riding problems: if k is small, the "informational" free-riding prevails and so Anarchy performs better. But, for large k, the "classical" free-riding problem becomes more important. The BD system is unaffected by "classical" free riding problems and so it will be preferred.

3.6. Extensions

In this section, I discuss two extensions of the model of binary public good provision. The goal is to inquire how robust is the result that the anarchic system outperforms the BD system. In the first extension, I allow the BD to use the services of a mediator that receives the messages from the agents and then makes his own report to the BD, who still makes all final decisions regarding transfers. In the second extension, I discuss the implications of the ex-post individual rationality constraints.

3.6.1. Mediator

The question I attempt to answer in this section is whether the presence of a mediator eliminates the advantage the anarchic system has over the centralized one when the public good is binary. I define the mediator to be someone that receives the messages sent by the agents and transmits its own message to the BD from an arbitrarily large message set that WLOG can still be M. This means that the mediator can pool the information gathered from the agents when communicating with the BD. I assume the mediator is indifferent among all outcomes so that no additional incentive constraints are required.

Let $\zeta = \{\zeta^m : m \in M\}$ denote a generic strategy of the mediator where $\zeta^m \in \nabla(M)$ - the mediator selects a probability distribution over the set of messages M, for every message m he has received. A mediator equilibrium is a strategy profile (σ, ζ, ξ) such that, just like

in the BD equilibrium, the agents and the BD choose their actions in order to maximize their expected utility and beliefs are updated according to Bayes' rule whenever possible. No behavior is imposed on the mediator.[14]

The first result is that the presence of a mediator does make a difference.

Proposition 30 *There is a mediator equilibrium that has a strictly higher expected welfare than the value of the BD system as long as* $\left(\overline{\rho},\overline{t}\right) \neq (\rho^*, t^*)$.

Proof. See appendix. ■

The weak part of the proposition is straightforward. We can simply have a passive mediator and replicate the BD equilibrium. The strict part comes from the fact that the mediator can use all the information provided by the agents when communicating with the BD. We illustrate the point with $N = 2$. Recall that, from proposition 30, we have that the highest expected welfare BD equilibrium was defined by $\left(\widetilde{\sigma}(s), \widetilde{\xi}\right)$ for some $s \in (0,1)^2$. If the first best allocation (ρ^*, t^*) is not incentive compatible, it must be that such that $s \neq (0,0)$. In this mediator setting, we can replicate the exact same equilibrium by having agents report truthfully their types to the mediator (so that an H (L) message is as if the agent says he is of type $\overline{v}$ $(\underline{v})$) and then allowing the mediator to mix between the messages to be sent to the BD as follows:

	HH	HL	LH	LL
$(\overline{v},\overline{v})$	$(1-s_1)(1-s_2)$	$(1-s_1)s_2$	$(1-s_2)s_1$	$s_1 s_2$
$(\overline{v},\underline{v})$	0	$(1-s_1)$	0	s_1
$(\underline{v},\overline{v})$	0	0	$(1-s_2)$	s_2
$(\underline{v},\underline{v})$	0	0	0	1

where the rows represent the messages received by the mediator, which are then mapped to the messages in the columns that are received by the BD. Assume that $\widehat{i} = 1$, so that the BD provides the public good if and only if the message he receives contains at least one H.

[14]Notice that the assumption that the mediator has no strict preference over any outcome increases the set of possible mediator equilibria.

This means that after LL the public good does not get provided. Then, instead of using the previous strategy, we can have the mediator send the following:

	HH	HL	LH	LL
$(\overline{v},\overline{v})$	$(1-s_1)(1-s_2)+\mathbf{s_1 s_2}$	$(1-s_1)\,s_2$	$(1-s_2)\,s_1$	$\mathbf{0}$
$(\overline{v},\underline{v})$	0	$(1-s_1)$	0	s_1
$(\underline{v},\overline{v})$	0	0	$(1-s_2)$	s_2
$(\underline{v},\underline{v})$	0	0	0	1

We can have the mediator send a HH message when he knows both agents are of type $\overline{v}$ rather than sending LL. The payoffs after any message received by the BD do not change. The only thing that changes is that the expected welfare of a high type agent is now increased, provided he reports truthfully, as the public good is being provided more often. Low valuation agents' incentives to misreport are even smaller because after HH the good is provided but they would have to contribute $c > \underline{v}$. Hence, the fact that the mediator can use all the information from both agents when communicating with the BD does makes a difference and allows for a higher expected welfare.

The next result states that, even though the presence of the mediator strictly improves upon the value of the BD system, it is still strictly worse than the anarchic alternative.

Proposition 31 *If* $\left(\overline{\rho},\overline{t}\right) \neq (\rho^*,t^*)$*, there is no mediator equilibrium that can implement* $\left(\overline{\rho},\overline{t}\right)$.

Proof. Notice that, WLOG, we can restrict our attention to truthful reporting by the agents. This is because it is always possible to rewrite any other type of reporting profile that is a part of a mediator equilibrium by having the mixing over the messages being done by the mediator, rather than by the agents, as I showed in the proof of proposition 36. Assuming truthful reporting by the agents, consider the mediator's choice. In order to implement $\left(\overline{\rho},\overline{t}\right)$ it must be that the mediator fully reveals the agents' types to the BD by the same argument as in proposition 28. But it that was the case, high valuation agents would prefer to misreport given that $\left(\overline{\rho},\overline{t}\right) \neq (\rho^*,t^*)$. ■

The mediator's presence does not solve the informational free riding problem because it is still not possible to use all the information gathered by the agents. Even with the presence of the mediator, the agents know the BD will make all the final decisions, which prevents them from communicating as much as they would have if the BD did not exist.

3.6.2. Individually Rationality Constraints

In my exposition so far I have assumed that the BD was not allowed to make any agent worse off as a result of the provision of the public good. In this section, I discuss the reasons such assumption is made, and the consequences of relaxing it.

The main argument for that assumption is that, if the BD imposes a loss on an agent, the agent would have an incentive to simply not participate in the mechanism, which would be somewhat equivalent to him leaving the community and looking for a different one where the public good is not provided. Green and Laffont (1979, p.121) defend the assumption by arguing that it represents the "ethical precept that no one should have to be forced to participate in the mechanism, and each has a right to withdraw from the system, abstain from consuming any public goods, and live independently with his endowment intact".

In this paper's context, it is also possible to interpret the ex-post individual rationality constraint as being part of the social welfare function. In particular, such constraint is equivalent to having a strongly negative welfare in case one of the agents has a negative utility - a rawlsian welfare function, for example, would have that property.

The imposition of this constraint only plays a role in the centralized system. In Anarchy, the constraint must be satisfied by default given that every agent always has the opportunity to simply not contribute, which guarantees that he cannot be made worse off. However, it may be in the best interest of the BD to have some agent have a negative payoff when he is unsure of the valuation of that agent. For example, if the BD believes the agent's type is high with a probability very close to 1, he will be very inclined to impose a high transfer on such agent which, in the unlikely event that the agent ends up ends up having a low valuation, may lead him to have a negative payoff. In this sense, the ability to impose

negative utilities on the agents will be an extra tool the BD has that an anarchic system cannot match.

The impact of the elimination of this constraint will depend very much on how heavily penalized by the welfare function are negative individual utilities. If society (and the BD) are relatively unfazed by the fact that agents are having a negative payoff it may be the case that the centralized system will outperform the anarchic one. Consider, for example, the case for which N is large. We know from Mailath and Postlewaite (1990) that any incentive compatible and individually rational mechanism will only be able to provide the public good with a very small probability, so the expected welfare from Anarchy will be close to 0. However, for a large N, the BD will know whether the good should be provided or not due to the law of large numbers. As long as π (the prior probability that $v_n = \overline{v}$ for all n) is large the BD could simply impose that the public good is to be provided by having all agents make a transfer of c, without soliciting any information. This means that the removal of the individual rationality constraint may make a BD system more appealing than the anarchic one.

3.7. Conclusion

In this paper, I have compared two systems of provision of public goods: a centralized system where the BD has the ultimate power to decide transfers; and an anarchic system where agents are free to communicate and select their own transfers. The main result is that the preferred system depends on how important is the "informational" free riding problem relative to the "classical" free riding problem, which is measured by the parameter k in this analysis. If k is small, and the "informational" free riding problem dominates, then the anarchic system is preferred; but as k grows and the "classical" free riding problem becomes more important, the BD system will dominate.

One aspect of this analysis that is interesting is that, when we think of the BD as lacking commitment power, it may be better to have a non-benevolent dictator. The problem that the absence of commitment power brings to the BD is that he cannot help but to take his

preferred action whenever the opportunity presents itself. So, if, for example, the dictator is indifferent among all alternatives available, he will not have such problems and will be able to act as if he had commitment power. Of course, thinking of a government that is indifferent to all alternatives does not seem to be very reasonable. However, there are other more reasonable preferences that could allow for a better outcome. Say that the dictator only cares about efficiency, even though the welfare function is still inequality averse. In that case, the dictator is able to commit, at least with respect to the transfers in the event that the public good is provided, which would allow him to extract more information from the agents.

The idea I wish to convey is that, if we are to have centralized provision of public goods, it may be desirable to have a government who does not feel the need to make the high valuation agents the ones that pay for most of the good. The problem is how that would happen. What political structure would have to be in place so that agents with these specific preferences would end up the rulers? The answer to this question does not seem trivial as the purpose of government is not exhausted by the administration of public goods and inequality aversion seems to be an important property of a successful government particularly in issues like wealth redistribution.

Finally, for future research, it would be interesting to investigate to what extent it is possible to construct other systems of public good provision that do not rely on commitment power. One possibility is to explore the idea of allowing agents to delegate their transfer decisions to representatives. Even though a more thorough analysis is required, it seems that this system would be able to go around the individual rationality constraints that limit the effectiveness of the anarchic system. For example, in a system with N representatives, as long π is large, it would always be possible to provide the public good fully by having agents sending uninformative messages to their representatives, which, after communicating with all other representatives, would then decide the agent they represent should contribute c, because they believe the agent they represent is likely to have a high valuation. Such outcome would not be possible in the anarchic system as low valuation agents would prefer

not to make such contribution. A system that incorporates representatives could then preserve the idea that decentralization leads to more information sharing while, at the same time, be less constrained by the need to have positive utilities for all agents, regardless of their type.

APPENDIX

A.1. Appendix to Chapter 1

A.1.1. Proof of Proposition 1

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

$$\sum_{t_{-n} \in T_{-n}} \pi(g, t_{-n}, \theta) \geq \alpha \sum_{t_{-n} \in T_{-n}} \pi(i, t_{-n}, \theta) \Leftrightarrow$$

$$\sum_{t_{-n} \in T_{-n}} \pi(g, t_{-n}) \pi(\theta | g, t_{-n}) \geq \alpha \sum_{t_{-n} \in T_{-n}} \pi(i, t_{-n}) \pi(\theta | i, t_{-n}) \Leftrightarrow$$

$$l(\theta_n) \geq \alpha \frac{\sum_{t_{-n} \in T_{-n}} \pi(i, t_{-n}) \prod_{\widetilde{n} \neq n} \pi(\theta_{\widetilde{n}} | t_{\widetilde{n}})}{\sum_{t_{-n} \in T_{-n}} \pi(g, t_{-n}) \prod_{\widetilde{n} \neq n} \pi(\theta_{\widetilde{n}} | t_{\widetilde{n}})} \Leftrightarrow$$

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

A.1.2. 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

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

subject to

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

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 $\theta \in \Theta$,

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

which can be written as

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

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

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

where ties are resolved in favor of an acquittal. 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 $B_n\left(\overline{\phi}_n, 0\right) = 1$.

A.1.3. 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,

$$B_n^g\left(x_n^{SB}\right) = \phi \sum_{t_{-n} \in T_{-n}} \frac{\pi(g, t_{-n})}{\pi(t_n = g)} \int_{\theta_n^{SB}(t_{-n})}^{1} \pi(\theta_n | t_n = g)\, d\theta_n = 1 \tag{A.2}$$

and

$$B_n^i\left(x_n^{SB}\right) = \phi \sum_{t_{-n} \in T_{-n}} \frac{\pi(i, t_{-n})}{\pi(t_n = i)} \int_{\theta_n^{SB}(t_{-n})}^{1} \pi(\theta_n | t_n = i)\, d\theta_n \tag{A.3}$$

Given (A.2) we have that (A.3) 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 $\lim_{\phi\to\infty}\ \theta_n^{SB}\left(t_{-n}\right) = 1$ for all $t_{-n} \in T_{-n}$ (for otherwise the expected punishments would become arbitrarily large, violating the constraints), which implies that $\lim_{\phi\to\infty}\ l\left(\theta_n^{SB}\left(t_{-n}\right)\right) = \infty$, and so $\lim_{\phi\to\infty}\ B_n^i\left(x_n^{SB}\right) = 0$ for all n.

A.1.4. Proof of Lemma 8

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

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

The goal of the proof is to show that it is possible to eliminate one such message. In this way, the set of messages only needs to be large enough as $\mathbb{R}_+ \cup \{c\}$ because the range of $r_n(\cdot)$ is $\mathbb{R}_+$ to which one adds the confessing message c.

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

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) $\overline{x}\left(m'_n, m_{-n}, \theta\right) = \left(\begin{array}{c} \frac{\sigma_n(t_n, m'_n)}{\sigma_n(t_n, m'_n)+\sigma_n(t_n, m''_n)} x\left(m'_n, m_{-n}, \theta\right) \\ + \frac{\sigma_n(t_n, m''_n)}{\sigma_n(t_n, m'_n)+\sigma_n(t_n, m''_n)} x\left(m''_n, m_{-n}, \theta\right) \end{array} \right)$ for $t_n = i, g$,

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. I want to show that the new system $(\overline{x}, \overline{\sigma})$ is still incentive compatible, renegotiation proof and leaves the expected utility of the principal unchanged.

Notice that

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

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 $\pi^{\overline{\sigma}}\left(m_{n}^{\prime},m_{-\widehat{n},n},\theta|t_{\widehat{n}}\right)$ is equal to

$$\sum_{t_{n}\in\{i,g\}}\left[\overline{\sigma}_{n}\left(t_{n},m_{n}^{\prime}\right)\pi\left(\theta_{n}|t_{n}\right)\pi\left(\theta_{\widehat{n}}|t_{\widehat{n}}\right)\sum_{t_{-\widehat{n},n}}\pi\left(t_{\widehat{n}},t_{n},t_{-\widehat{n},n}|t_{\widehat{n}}\right)\prod_{\widetilde{n}\neq n,\widehat{n}}\pi\left(\theta_{\widetilde{n}}|t_{\widetilde{n}}\right)\sigma_{\widetilde{n}}\left(m_{\widetilde{n}},t_{\widetilde{n}}\right)\right]$$

Given that

$$\begin{aligned}&\pi^{\overline{\sigma}}\left(m_{n}^{\prime},m_{-\widehat{n},n},\theta|t_{\widehat{n}}\right)\overline{x}_{n}\left(m_{\widehat{n}},m_{n}^{\prime},m_{-\widehat{n},n},\theta\right)+\pi^{\overline{\sigma}}\left(m_{n}^{\prime\prime},m_{-\widehat{n},n},\theta|t_{\widehat{n}}\right)\overline{x}_{n}\left(m_{\widehat{n}},m_{n}^{\prime\prime},m_{-\widehat{n},n},\theta\right)\\&=\pi^{\sigma}\left(m_{n}^{\prime},m_{-\widehat{n},n},\theta|t_{\widehat{n}}\right)x_{n}\left(m_{\widehat{n}},m_{n}^{\prime},m_{-\widehat{n},n},\theta\right)+\pi^{\sigma}\left(m_{n}^{\prime\prime},m_{-\widehat{n},n},\theta|t_{\widehat{n}}\right)x_{n}\left(m_{\widehat{n}},m_{n}^{\prime\prime},m_{-\widehat{n},n},\theta\right)\end{aligned}$$

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 $m_{n}^{\prime\prime}$ is not strictly preferred to any other message and the expected punishment of sending any other message remains unchanged. It is also renegotiation proof because, for all m_{-n},θ and for all $\widehat{n}$ (including n)

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

A.1.5. 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 or equal to that of the guilty agent.

Notice that it is possible to write

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

Define $E_n \equiv \left\{(m_{-n},\theta) \in M_{-n} \times [0,1]^N : x_n^\sigma\left(m_n^\sigma, m_{-n}, \theta\right) = 1\right\}$. If $E_n = \varnothing$ or $\rceil E_n = \varnothing$ then the expected punishment of the agent when sending message m_n^σ is independent of his type. If $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)} < 1$ for all $e_n \in E_n$ then $\int\limits_{e_n \in E_n} \pi\left(e_n|t_n=g\right) de_n > \int\limits_{e_n \in E_n} \pi\left(e_n|t_n=i\right) de_n$ and so the expected punishment of the agent when sending message m_n^σ is higher if he is guilty. Finally, if there is $e_n' \in E_n$ such that $\frac{\pi(e_n'|t_n=i)}{\pi(e_n'|t_n=g)} \geq 1$ and given that $x_n^\sigma\left(m_n^\sigma, m_{-n}, \theta\right)$ is decreasing with $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)}$, then it must be that $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)} > 1$ for all $e_n \notin E_n$. Hence, $\int\limits_{e_n \notin E_n} \pi\left(e_n|t_n=g\right) de_n < \int\limits_{e_n \notin E_n} \pi\left(e_n|t_n=i\right) de_n$ which implies that $\int\limits_{e_n \in E_n} \pi\left(e_n|t_n=g\right) de_n > \int\limits_{e_n \in E_n} \pi\left(e_n|t_n=i\right) de_n$ and so, 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 $r_n\left(m_n^\sigma\right)$, except for message c, where the agent is always believed to be guilty with certainty. Hence, it follows that the system is renegotiation proof because $\gamma_n^\sigma\left(m_n^\sigma, m_{-n}, \theta\right) \leq \gamma_n^\sigma\left(m_n, m_{-n}, \theta\right)$ for all $m_n \in \mathbb{R}_+$.

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

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

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

$$\kappa_n^\sigma\left(m, \theta, x'_n(m, \theta)\right) = \left(\pi^\sigma(t_n = g|m, \theta) - \pi^\sigma(t_n = i|m, \theta)\right) x'_n(m, \theta)$$

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 message m_n^σ 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 systems, so 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

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

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

A.1.6. Proof of Proposition 10

Take any optimal strategy profile $\widetilde{\sigma}$ and suppose the statement is false: system $\left(x^{\widetilde{\sigma}}, \widetilde{\sigma}\right)$ is not a CIS. 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 $m_1' \notin \left\{c, m_1^{\widetilde{\sigma}}\right\}$. In particular, assume that $r_1\left(m_1^{\widetilde{\sigma}}\right) < r_1(m_1') < \infty$ 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 $\widehat{\sigma}_1(g, m_1') = \widetilde{\sigma}_1(g, m_1') - v$, $\widehat{\sigma}_1(g, c) = \widetilde{\sigma}_1(g, c) + v$, where v is such that

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

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

Write $\overline{V}(\sigma) \equiv \widehat{V}(x^{\sigma}, \sigma)$ for all σ. Notice that $\overline{V}_1(\widehat{\sigma}) = \overline{V}_1(\widetilde{\sigma})$. 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 B is equal to

$$\begin{aligned}
&v\left(\begin{array}{c}\pi\left(g,g,t_{-1,n}\right)\widetilde{\sigma}_n\left(g,m_n\right)\pi\left(\theta_n|t_n=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{array}\right)\gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}},c,m_{-1,n}\right),\theta\right)+\\
&\left[\begin{array}{c}\left(\begin{array}{c}\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'\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'\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'\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'\right)\pi\left(\theta_1|t_1=i\right)\end{array}\right)\\ *\gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}},m_1^{\widetilde{\sigma}},m_{-1,n}\right),\theta\right)\end{array}\right]\\
&-\left[\begin{array}{c}\left(\begin{array}{c}\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'\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'\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'\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'\right)\pi\left(\theta_1|t_1=i\right)\end{array}\right)\\ *\gamma_n^{\widetilde{\sigma}}\left(\left(m_n^{\widetilde{\sigma}},m_1',m_{-1,n}\right),\theta\right)\end{array}\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

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

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.[1]

A.1.7. Proof of Proposition 11

In a CIS, only two messages are sent: c and $\overline{c}$ 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 $\sigma_n^{CIS}(g,c) = \tau_n$ for all n. Also, let $\overline{V}(\tau)$ denote the corresponding expected utility of the principal. A trial system is characterized by $\tau = \underline{\tau} \equiv (0,...,0)$.

I show the statement by showing that

$$\frac{\partial \overline{V}_n}{\partial \tau_n}(\underline{\tau}) = 0 \text{ for all } n \tag{A.4}$$

and

$$\frac{\partial \overline{V}_{\widehat{n}}}{\partial \tau_n}(\underline{\tau}) \geq 0 \text{ for all } n \text{ and } \widehat{n} \tag{A.5}$$

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 $\overline{V}(\tau) = \sum_{n=1}^{N} \overline{V}_n(\tau)$ where

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

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

where

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

and

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

The threshold $\theta_n^{CIS}\left(m_{-n},\theta_{-n}\right)$ is such that

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

As for (A.4), notice that

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

Given that, when $\tau_n=0$,

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

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

Now, consider (A.5). Notice that one can write $\overline{V}_{\widehat{n}}(\tau)$ as

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

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

$$\int_{m_{-\widehat{n},n},\theta_{-\widehat{n}}} \left[\begin{array}{l} -A^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right),\theta_{-\widehat{n}}\right)* \\ \qquad \frac{d\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}{d\tau_n} \\ -A^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{\widehat{n}}^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{-\widehat{n}}\right),\theta_{-\widehat{n}}\right)* \\ \qquad \frac{d\theta_{\widehat{n}}^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}{d\tau_n} \end{array} \right] d\theta_{-\widehat{n}}dm_{-\widehat{n},n} \tag{A.6}$$

$$+ \int_{m_{-\widehat{n},n},\theta_{-\widehat{n}}} \left[\begin{array}{l} \int_{\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}^{1} \frac{dA^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right)}{d\tau_n}d\theta_{\widehat{n}} \\ + \int_{\theta_{\widehat{n}}^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}^{1} \frac{dA^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right)}{d\tau_n}d\theta_{\widehat{n}} \end{array} \right] d\theta_{-\widehat{n}}dm_{-\widehat{n},n} \tag{A.7}$$

Notice that (A.6) is equal to 0 when $\tau_{\widehat{n}}=0$ given that

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

Let

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

and

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

Then, condition (A.7) can be written as

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

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

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

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

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

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

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

A.1.8. 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

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

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

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

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

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

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

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

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

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

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 $\theta_{-n} \in \Theta_{-n}$,

$$\lim_{\theta_n \to 0} x_n^{Tr}((\theta_n, \theta_{-n})) = 0$$

and given that $\lim_{\theta_n \to 1} l(\theta_n) = \infty$ it must be that, for all $\theta_{-n} \in \Theta_{-n}$,

$$\lim_{\theta_n \to 1} x_n^{Tr}((\theta_n, \theta_{-n})) = 1.$$

A.1.9. Proof of Proposition 15

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

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

subject

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

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

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

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

$$\{x_n(g, t_{-n}, \theta) \geq 0\}$$

and

$$\{x_n(g, t_{-n}, \theta) \leq \phi\}.$$

Given that

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

and

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

and

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

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

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

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

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

because the incentive constraint binds.

A.1.10. Proof of Proposition 16

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

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

which can be written as

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

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

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

Notice that

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

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

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

This shows i).

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

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

which implies that

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

Therefore, ii) follows.

The variable λ_n is such that

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

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

A.1.11. 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 $\widehat{t}_n\neq i$,

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

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

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

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

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

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

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

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

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

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

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

A.1.12. 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 18.1 *For all n, there is $\left(\beta_n^g, \beta_n^i\right) \in [0,1]^N$ such that either*

A) for all (t_n, β_n),

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

or B) for all (t_n, β_n),

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

Proof of Lemma 18.1 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

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

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 $\beta_n\in\Gamma_=^g$,

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

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

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

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

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

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

where $\zeta^{\overline{c}}_n(m_{-n}, \theta) \geq 0$ and $\eta^{\overline{c}}_n(m_{-n}, \theta) \geq 0$ 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:

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

where

$$h(\theta_n) = \frac{\int_{\beta_n \in \Gamma^g_{=} \cup \Gamma^g_{\overline{c}}} \pi(\beta_n | t_n = g)\, \lambda^g(\beta_n)\, \pi(\theta_n | \beta_n)\, d\beta_n}{\int_{\beta_n \in \Gamma^i_{=} \cup \Gamma^i_{\overline{c}}} \pi(\beta_n | t_n = i)\, \lambda^i(\beta_n)\, \pi(\theta_n | \beta_n)\, d\beta_n}$$

and

$$k(m_{-n}, \theta_{-n}) = \frac{\int_{\beta_{-n}} \sum_{t_{-n}} \pi(g, t_{-n})\, \pi(\beta_{-n} | t_{-n})\, \pi(\theta_{-n} | \beta_{-n})\, \pi(m_{-n} | t_{-n}, \beta_{-n})\, d\beta_{-n}}{\alpha \int_{\beta_{-n}} \sum_{t_{-n}} \pi(i, t_{-n})\, \pi(\beta_{-n} | t_{-n})\, \pi(\theta_{-n} | \beta_{-n})\, \pi(m_{-n} | t_{-n}, \beta_{-n})\, d\beta_{-n}}$$

Notice that

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

where

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

and

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

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

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

Lemma 18.2 For all n, there is $\left(\beta_n^g, \beta_n^i\right) \in [0,1]^N$ such that for all (t_n, β_n),

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

Proof of Lemma 18.2 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 are only two pairs $\left(i, \beta_n^i\right)$ 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

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

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

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

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

A.1.13. 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 $x_y\left(t, \theta\right)$ 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 $\pi^{y_\theta}\left(t, \theta | t_n\right)$. 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',

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

Given that the previous expression holds for all $\widehat{\theta}$, it follows that, for all t'_n,

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

where $\pi\left(\widehat{\theta}|t_n\right)$ refers to the density of $\widehat{\theta}$ conditional of the agent's type t_n. Now, I want to group into disjoint sets the evidence that, given the strategy of the principal, induces the same posterior on the agent. More formally denote by $\chi_{\widehat{\theta}}\equiv\left\{\theta\in\Theta:y_\theta=y_{\widehat{\theta}}\right\}$ and $\widehat{\Theta}\equiv\left\{\widehat{\theta}\in\Theta:\text{for all }\theta\text{ such that }\pi_{\widehat{\theta}}=\pi_\theta\text{ then }\widehat{\theta}\prec_l\theta\right\}$ where $\prec_l$ denotes the lexicographic ordering[2]. 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

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

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

A.1.14. 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

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

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, $\widehat{y}_{\widehat{\theta},n}\left(i,\left(g,..,g\right),\left(0,..,0\right)\right) = \widehat{y}_{\theta,n}\left(g,\left(g,..,g\right),\left(0,..,0\right)\right)$. As for $\widehat{y}_{\widehat{\theta},n}\left(t,\theta\right)$ 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

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

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.

A.2. Appendix to Chapter 2

A.2.1. Proof of Proposition 21

By Bayesian updating the agent's beliefs, it follows that $y|s_1, s_2 \sim N\left(\widehat{\mu}_y, \widehat{\sigma}_y^2\right)$, where

$$\widehat{\mu}_y = \sigma_y^2 \frac{s_2 - h\mu_x + \left(h^2\sigma_x^2 + 1\right) s_1}{\sigma_y^2 + \left(h^2\sigma_x^2 + 1\right)\left(\sigma_y^2 + 1\right)}$$

and

$$\widehat{\sigma}_y^2 = \frac{\left(h^2\sigma_x^2 + 1\right)\sigma_y^2}{\sigma_y^2 + \left(h^2\sigma_x^2 + 1\right)\left(\sigma_y^2 + 1\right)}$$

Therefore,

$$\Pr\left\{\widehat{\mu}_y \geq 0\right\} = \Pr\left\{q \geq 0\right\} = \int_{-\infty}^{\infty} f_{s_1}\left(\widetilde{s}_1\right) \Pr\left\{q \geq 0 | s_1 = \widetilde{s}_1\right\} d\widetilde{s}_1$$

where

$$q\left(s_1, s_2\right) = s_2 - h\mu_x + \left(h^2\sigma_x^2 + 1\right) s_1$$

Given that s_1 is symmetric distributed around 0 it follows that

$$\Pr\left\{\widehat{\mu}_y \geq 0\right\} = \int_{0}^{\infty} f_{s_1}\left(\widetilde{s}_1\right) \left[\begin{array}{c} \Pr\left\{q\left(s_1, s_2\right) \geq 0 | s_1 = \widetilde{s}_1\right\} \\ + \Pr\left\{q\left(s_1, s_2\right) \geq 0 | s_1 = -\widetilde{s}_1\right\} \end{array} \right] d\widetilde{s}_1$$

Notice that, because $y|s_1$, $x|s_1$ and $\varepsilon_2|s_1$ are independent and normally distributed, it follows that $s_2|s_1$ is also normally distributed, which implies that

$$q|s_1 \sim N\left(\left(\frac{\sigma_y^2}{\sigma_y^2 + 1} + h^2\sigma_x^2 + 1\right) s_1, \frac{\sigma_y^2}{\sigma_y^2 + 1} + h^2\sigma_x^2 + 1\right)$$

Therefore,

$$\begin{aligned}&\Pr\{q \geq 0 | s_1 = \widetilde{s}_1\} + \Pr\{q \geq 0 | s_1 = -\widetilde{s}_1\} \\ &= 2 - \Phi\left(-\widetilde{s}_1 \sqrt{\frac{\sigma_y^2}{\sigma_y^2+1} + \Psi^2(\widetilde{s}_1)\sigma_x^2 + 1}\right) - \Phi\left(\widetilde{s}_1 \sqrt{\frac{\sigma_y^2}{\sigma_y^2+1} + \Psi^2(-\widetilde{s}_1)\sigma_x^2 + 1}\right)\end{aligned}$$

Given that

$$\sqrt{\frac{\sigma_y^2}{\sigma_y^2+1} + h^2\sigma_x^2 + 1}$$

is strictly increasing with h and Ψ is increasing it follows that

$$\Phi\left(-\widetilde{s}_1 \sqrt{\frac{\sigma_y^2}{\sigma_y^2+1} + \Psi^2(\widetilde{s}_1)\sigma_x^2 + 1}\right) + \Phi\left(\widetilde{s}_1 \sqrt{\frac{\sigma_y^2}{\sigma_y^2+1} + \Psi^2(-\widetilde{s}_1)\sigma_x^2 + 1}\right) \leq 1$$

for all $\widetilde{s}_1 \geq 0$, which proves the result. If h is not constant, it follows that the above inequality will be strict for all $\widetilde{s}_1 \geq \overline{s}_1$ for some $\overline{s}_1$, which implies that $\Pr\{\widehat{\mu}_y \geq 0\} > \frac{1}{2}$.

A.2.2. Proof of Proposition 22

Notice that

$$\widetilde{\mu}_y = \delta_y^2 \frac{s_2 - h\mu_x + \left(h^2\sigma_x^2 + 1\right)s_1}{\delta_y^2 + \left(h^2\sigma_x^2 + 1\right)\left(\delta_y^2 + 1\right)}$$

and so

$$\widetilde{o} = \delta_y^2 \frac{s_2 - h\mu_x + \left(h^2\sigma_x^2 + 1\right)s_1}{\delta_y^2 + \left(h^2\sigma_x^2 + 1\right)\left(\delta_y^2 + 1\right)} - y$$

I start by showing the first part of the result. Notice that it is enough to show that

$$E\left(\widetilde{\mu}_y\right) \geq 0$$

if Ψ is increasing, with the inequality being strict if Ψ is strictly increasing.

Notice that

$$E\left(\widetilde{\mu}_y\right) = E_{s_1}\left(E\left(\widetilde{\mu}_y|s_1\right)\right)$$

where

$$E\left(\widetilde{\mu}_y|s_1\right) = d\left(h\right)s_1$$

and

$$d\left(h\right) = \frac{\delta_y^2}{\sigma_y^2+1}\frac{\sigma_y^2+\left(h^2\sigma_x^2+1\right)\left(\sigma_y^2+1\right)}{\delta_y^2+\left(h^2\sigma_x^2+1\right)\left(\delta_y^2+1\right)}$$

Notice that $d\left(h\right)$ is strictly increasing with h if $\delta_y > \sigma_y$, strictly decreasing if $\delta_y < \sigma_y$ and constant if $\delta_y = \sigma_y$.

If $\delta_y > \sigma_y$ and Ψ is increasing, it follows that

$$\begin{aligned} E\left(d\left(\Psi\left(s_1\right)\right)s_1\right) &= \int_{-\infty}^{\infty} f_{s_1}\left(\widetilde{s}_1\right) d\left(\Psi\left(\widetilde{s}_1\right)\right)\widetilde{s}_1 d\widetilde{s}_1 \\ &= \int_{0}^{\infty} f_{s_1}\left(\widetilde{s}_1\right)\left(d\left(\Psi\left(\widetilde{s}_1\right)\right) - d\left(\Psi\left(-\widetilde{s}_1\right)\right)\right)\widetilde{s}_1 d\widetilde{s}_1 \\ &\geq 0 \end{aligned}$$

where the inequality is strict if Ψ is strictly increasing.

As for the second part, notice that one can write

$$\Pr\left\{\widetilde{o} \geq 0\right\} = \Pr\left\{p \geq 0\right\}$$

where

$$p = \delta_y^2 h\left(x-\mu_x\right) + \delta_y^2\varepsilon_2 + \delta_y^2\left(h^2\sigma_x^2+1\right)s_1 - \left(h^2\sigma_x^2+1\right)\left(\delta_y^2+1\right)y$$

Notice that

$$\Pr\left\{p \geq 0\right\} = \int_{0}^{\infty} f_{s_1}\left(\widetilde{s}_1\right)\left[\begin{array}{c} \Pr\left\{p \geq 0|s_1 = \widetilde{s}_1\right\} \\ +\Pr\left\{p \geq 0|s_1 = -\widetilde{s}_1\right\} \end{array}\right] d\widetilde{s}_1$$

Because $x|s_1$, $y|s_1$ and $\varepsilon_2|s_1$ are all independent and normally distributed, it follows that

$$p|s_1 \sim N\left(\mu_p\left(h^2\right)s_1, \sigma_p^2\left(h^2\right)\right)$$

where

$$\mu_p\left(h^2\right) = \left(h^2\sigma_x^2+1\right)\left(\delta_y^2 - \sigma_y^2\frac{\delta_y^2+1}{\sigma_y^2+1}\right)$$

and

$$\sigma_p^2\left(h^2\right) = \delta_y^4 h^2\sigma_x^2 + \delta_y^4 + \left(h^2\sigma_x^2+1\right)^2\left(\delta_y^2+1\right)^2\frac{\sigma_y^2}{\sigma_y^2+1}$$

Therefore,

$$\Pr\left\{p \geq 0|s_1 = \widetilde{s}_1\right\} + \Pr\left\{p \geq 0|s_1 = -\widetilde{s}_1\right\} = 2 - \Phi\left(-\frac{\mu_p\left(\Psi^2\left(s_1\right)\right)}{\sigma_p\left(\Psi^2\left(s_1\right)\right)}\widetilde{s}_1\right) - \Phi\left(\frac{\mu_p\left(\Psi^2\left(-s_1\right)\right)}{\sigma_p\left(\Psi^2\left(-s_1\right)\right)}\widetilde{s}_1\right)$$

Notice that if $\delta_y > \sigma_y$, it follows that $\frac{\mu_p(h)}{\sigma_p(h)}$is strictly increasing with h, if $\delta_y < \sigma_y$ it is strictly decreasing and if $\delta_y = \sigma_y$ it is constant and equal to 0.

Therefore, if Ψ is increasing and $\delta_y > \sigma_y$, it is the case that

$$\Phi\left(-\frac{\mu_p\left(\Psi^2\left(s_1\right)\right)}{\sigma_p\left(\Psi^2\left(s_1\right)\right)}\widetilde{s}_1\right) + \Phi\left(\frac{\mu_p\left(\Psi^2\left(-s_1\right)\right)}{\sigma_p\left(\Psi^2\left(-s_1\right)\right)}\widetilde{s}_1\right) \leq 1$$

for all $\widetilde{s}_1 > 0$, which shows the result. It is also clear that if Ψ is not a constant there will be some $\overline{s}_1 > 0$ such that the previous inequality holds for all $\widetilde{s}_1 \geq \overline{s}_1$.

A.2.3. Proof of Proposition 23

The problem of the adviser is to select $\Psi\left(s_1\right)$ for all s_1 such that

$$\Psi\left(s_1\right) \in \arg\max_{h\in\left[0,\overline{h}\right]} E\left(g\left(y+hx+\varepsilon_2\right)|s_1\right) - ch$$

Notice that it must be that, for all s_1, $\Psi(s_1) \in \left\{0, \overline{h}\right\}$ because, if not, it would have to be that

$$E\left(xg'\left(y + \Psi(s_1)\, x + \varepsilon_1\right)|s_1\right) = c$$

and

$$E\left(x^2 g''\left(y + \Psi(s_1)\, x + \varepsilon_1\right)|s_1\right) \leq 0$$

which is not true because $g'' > 0$.

Therefore, it follows that $\Psi(s_1) = \overline{h}$ if

$$E\left(g\left(y + \overline{h}x + \varepsilon_2\right)|s_1\right) - E\left(g\left(y + \varepsilon_2\right)|s_1\right) \geq c\overline{h}$$

and $\Psi(s_1) = 0$ otherwise (where the assumption is that if there is a tie help is provided). The RHS is independent of s_1 so the proof is completed by showing the LHS is strictly increasing with s_1.

The LHS can be written as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\varepsilon_2}(\widetilde{\varepsilon}_2)\, f_{y|s_1}(\widetilde{y})\, f_x(\widetilde{x})\left(g\left(\widetilde{y} + \overline{h}\widetilde{x} + \widetilde{\varepsilon}_2\right) - g\left(\widetilde{y} + \widetilde{\varepsilon}_2\right)\right) d\widetilde{x}d\widetilde{y}d\widetilde{\varepsilon}_2$$

By doing a change of variables where $\widehat{y} = \widetilde{y} - \frac{\sigma^2}{\sigma^2+1}s_1$ and $\widehat{x} = \widetilde{x} - \mu_x$ it follows that the LHS is proportional to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\begin{array}{c} f_{\varepsilon_2}(\widetilde{\varepsilon}_2)\exp\left(-\frac{\widetilde{y}^2}{2\frac{\sigma_y^2}{\sigma_y^2+1}}\right)\exp\left(-\frac{\widetilde{x}^2}{2\sigma_x^2}\right) * \\ \left(g\left(\widetilde{y} + \frac{\sigma_y^2}{\sigma_y^2+1}s_1 + \overline{h}\widehat{x} + \overline{h}\mu_x + \varepsilon_2\right) - g\left(\widetilde{y} + \frac{\sigma_y^2}{\sigma_y^2+1}s_1 + \varepsilon_2\right)\right) \end{array}\right] d\widehat{x}d\widehat{y}d\widetilde{\varepsilon}_2$$

which is equal to

$$\int_{-\infty}^{\infty} f \int_{-\infty}^{\infty} \int_{0}^{\infty} \left[\begin{array}{c} f_{\varepsilon_2}(\widetilde{\varepsilon}_2) \exp\left(-\frac{\widetilde{y}^2}{2\frac{\sigma_y^2}{\sigma_y^2+1}}\right) \exp\left(-\frac{\widetilde{x}^2}{2\sigma_x^2}\right) * \\ \left(\begin{array}{c} g\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1+\overline{h}\widehat{x}+\overline{h}\mu_x+\varepsilon_2\right)-g\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1+\varepsilon_2\right) \\ +g\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1-\overline{h}\widehat{x}+\overline{h}\mu_x+\varepsilon_2\right)-g\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1+\varepsilon_2\right) \end{array} \right) \end{array} \right] d\widehat{x}d\widehat{y}d\widetilde{\varepsilon}_2$$

Therefore, the derivative of the previous expression with respect to s_1 is strictly positive if $g'' > 0$ because, for all $\widehat{x} > 0$, for all $\widehat{y}$ and $\widetilde{\varepsilon}_2$

$$\begin{aligned} & g'\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1+\overline{h}\widehat{x}+\overline{h}\mu_x+\varepsilon_2\right)+g'\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1-\overline{h}\widehat{x}+\overline{h}\mu_x+\varepsilon_2\right) \\ & > 2g'\left(\widetilde{y}+\frac{\sigma_y^2}{\sigma_y^2+1}s_1+\varepsilon_2\right) \end{aligned}$$

which concludes the proof.

A.2.4. Proof of Proposition 24

The parent's problem is to choose $h^n \geq 0$ for all $n = 1, ..., N$ such that $\sum_{n=1}^{N} h^n \leq \overline{h}$ in order to maximize $E\left[\sum_{n=1}^{N} g\left(s_2^n\right) \mid \{s_1^n\}_{n=1}^{N}\right]$.

First, I show that

$$\frac{\partial E\left[\sum_{n=1}^{N} g\left(s_2^n\right) \mid \{s_1^n\}_{n=1}^{N}\right]}{\partial h^{\widehat{n}}} > 0$$

for all $\widehat{n}$, which implies that $\sum_{n=1}^{N}\Psi^{n}=\overline{h}$. Notice that

$$\begin{aligned}\frac{\partial E\left[\sum_{n=1}^{N}g\left(s_{2}^{n}\right)\mid\left\{s_{1}^{n}\right\}_{n=1}^{N}\right]}{\partial h^{\widehat{n}}} &= E\left(x^{\widehat{n}}g'\left(y^{\widehat{n}}+h^{\widehat{n}}x^{\widehat{n}}+\varepsilon^{\widehat{n}}\right)|s_{1}^{\widehat{n}}\right)\\ &= E\left[E\left(x^{\widehat{n}}g'\left(y^{\widehat{n}}+h^{\widehat{n}}x^{\widehat{n}}+\varepsilon^{\widehat{n}}\right)|y^{\widehat{n}},\varepsilon^{\widehat{n}},s_{1}^{\widehat{n}}\right)|s_{1}^{\widehat{n}}\right]\end{aligned}$$

Notice that letting $\widetilde{x}^{\widehat{n}}=x^{\widehat{n}}-\mu_{x^{\widehat{n}}}$ it follows that

$$E\left(x^{\widehat{n}}g'\left(y^{\widehat{n}}+h^{\widehat{n}}x^{\widehat{n}}+\varepsilon^{\widehat{n}}\right)|y^{\widehat{n}},\varepsilon^{\widehat{n}},s_{1}^{\widehat{n}}\right)$$

is proportional to

$$\begin{aligned}&\int_{0}^{\infty}\left[\begin{array}{c}\exp\left(-\frac{\left(\widetilde{x}^{\widehat{n}}\right)^{2}}{\sigma_{x}^{2}}\right)*\\ \left(\begin{array}{c}\widetilde{x}^{\widehat{n}}g'\left(y^{\widehat{n}}+h^{\widehat{n}}\widetilde{x}^{\widehat{n}}+h^{\widehat{n}}\mu_{x^{\widehat{n}}}+\varepsilon^{\widehat{n}}\right)+\mu_{x^{\widehat{n}}}g'\left(y^{\widehat{n}}+h^{\widehat{n}}\widetilde{x}^{\widehat{n}}+h^{\widehat{n}}\mu_{x^{\widehat{n}}}+\varepsilon^{\widehat{n}}\right)\\ -\widetilde{x}^{\widehat{n}}g'\left(y^{\widehat{n}}-h^{\widehat{n}}\widetilde{x}^{\widehat{n}}+h^{\widehat{n}}\mu_{x^{\widehat{n}}}+\varepsilon^{\widehat{n}}\right)+\mu_{x^{\widehat{n}}}g'\left(y^{\widehat{n}}-h^{\widehat{n}}\widetilde{x}^{\widehat{n}}+h^{\widehat{n}}\mu_{x^{\widehat{n}}}+\varepsilon^{\widehat{n}}\right)\end{array}\right)\end{array}\right]d\widetilde{x}\\ &>\int_{0}^{\infty}\left[\begin{array}{c}\exp\left(-\frac{\left(\widetilde{x}^{\widehat{n}}\right)^{2}}{\sigma_{x}^{2}}\right)*\\ \widetilde{x}^{\widehat{n}}\left(g'\left(y^{\widehat{n}}+h^{\widehat{n}}\widetilde{x}^{\widehat{n}}+h^{\widehat{n}}\mu_{x^{\widehat{n}}}+\varepsilon^{\widehat{n}}\right)-g'\left(y^{\widehat{n}}-h^{\widehat{n}}\widetilde{x}^{\widehat{n}}+h^{\widehat{n}}\mu_{x^{\widehat{n}}}+\varepsilon^{\widehat{n}}\right)\right)\end{array}\right]\geq 0\end{aligned}$$

for all $y^{\widehat{n}},\varepsilon^{\widehat{n}},s_{1}^{\widehat{n}}$, where the last inequality follows from the fact that $g''>0$. This implies that

$$\frac{\partial E\left[\sum_{n=1}^{N}g\left(s_{2}^{n}\right)\mid\left\{s_{1}^{n}\right\}_{n=1}^{N}\right]}{\partial h^{\widehat{n}}}>0$$

and that $\sum_{n=1}^{N}\Psi^{n}=\overline{h}$.

Now I show that, for any arbitrary $\widehat{n}$, $\Psi^{\widehat{n}}\in\left\{0,\overline{h}\right\}$. Suppose not. Then there must be at least n', n'' such that $\Psi^{n'}\in(0,1)$ and $\Psi^{n''}\in(0,1)$. WLOG, say $n'=1$ and $n''=2$. Let

$\widetilde{h} = \overline{h} - \sum_{n=3}^{N} \Psi^n$. It follows that

$$\Psi^1 \in \arg \max_{h^1 \in [0,\widetilde{h}]} E\left[g\left(y^1 + h^1 x^1 + \varepsilon_2^1\right) | s_1^1\right] + E\left[g\left(y^2 + \widetilde{h} x^2 - h^1 x^2 + \varepsilon_2^2\right) | s_1^2\right]$$

Given that $\Psi^1 \notin \left\{0, \widetilde{h}\right\}$ it follows that

$$E\left[x^1 g'\left(y^1 + \Psi^1 x^1 + \varepsilon_2^1\right) | s_1^1\right] - E\left[x^2 g'\left(y^2 + \Psi^2 x^2 + \varepsilon_2^2\right) | s_1^2\right] = 0$$

and

$$E\left[\left(x^1\right)^2 g''\left(y^1 + \Psi^1 x^1 + \varepsilon_2^1\right) | s_1^1\right] + E\left[\left(x^2\right)^2 g''\left(y^2 + \Psi^2 x^2 + \varepsilon_2^2\right) | s_1^2\right] \leq 0$$

The second inequality is false given that $g'' > 0$ which is a contradiction. Therefore, $\Psi^{\widehat{n}} \in \left\{0, \overline{h}\right\}$.

Finally, given that, for all n, $\Psi^n \in \left\{0, \overline{h}\right\}$ and $\sum_{n=1}^{N} \Psi^n = \overline{h}$, it must be that the parents choose some $\widetilde{n}$ for which $\Psi^{\widetilde{n}} = \overline{h}$ such that

$$\widetilde{n} \in \arg \max_{n=1,\ldots,N} \left\{E\left[g\left(y^n + \overline{h} x^n + \varepsilon_2^n\right) | s_1^n\right] - E\left[g\left(y^n + \varepsilon_2^n\right) | s_1^n\right]\right\}$$

Therefore, to complete the proof, it suffices to show that, for all n,

$$\left\{E\left[g\left(y^n + \overline{h} x^n + \varepsilon_2^n\right) | s_1^n\right] - E\left[g\left(y^n + \varepsilon_2^n\right) | s_1^n\right]\right\}$$

is increasing with s_1^n, which I have on the proof of Proposition 23.

A.3. Appendix to Chapter 3

A.3.1. Proof of Proposition 27

To characterize the solution for the BD with commitment power problem I first solve a relaxed version of the same problem. Then, I show that $\left(\overline{\rho}, \overline{t}\right)$ is a solution of that relaxed problem and satisfies all constraints of the original problem, which implies that $\left(\overline{\rho}, \overline{t}\right)$ is also a solution of the original problem.

I start by summing all $\overline{v}$ and $\underline{v}$ incentive constraints which results in the following two conditions:

$$\overline{v} A(\rho) \geq B(\rho, t) \tag{A.8}$$

and

$$B(\rho, t) \geq \underline{v} A(\rho) \tag{A.9}$$

where

$$A(\rho) = \sum_{n=1}^{N} \sum_{v_{-n} \in \{\overline{v}, \underline{v}\}^{N-1}} \Pr\{v_{-n}\} \left(\rho\left(\overline{v}, v_{-n}\right) - \rho\left(\underline{v}, v_{-n}\right)\right)$$

and

$$B(\rho, t) = \sum_{n=1}^{N} \sum_{v_{-n} \in \{\overline{v}, \underline{v}\}^{N-1}} \Pr\{v_{-n}\} \left(\rho\left(\overline{v}, v_{-n}\right) t_n\left(\overline{v}, v_{-n}\right) - \rho\left(\underline{v}, v_{-n}\right) t_n\left(\underline{v}, v_{-n}\right)\right)$$

The relaxed problem that I consider is one where: a) the $\underline{v}$ incentive constraints are relaxed and b) the $\overline{v}$ incentive constraints are replaced by condition (A.8) - the relaxed $\overline{v}$ incentive constraint. The individual rationality constraints and the feasibility constraints are still part of this relaxed problem.

Consider first the case where (ρ^*, t^*) solves the relaxed problem. That implies that

condition (A.9) - the relaxed $\underline{v}$ incentive constraint - is also satisfied because $\rho^*(\overline{v}, v_{-n}) \geq \rho^*(\underline{v}, v_{-n})$ for all $v_{-n} \in \{\underline{v}, \overline{v}\}^{N-1}$ which implies that $A(\rho^*) \geq 0$. Given that (ρ^*, t^*) does not depend on the identity of the agents it follows that condition (3.2) holds. Conditions (3.3) and (3.4) also trivially hold. Then, it must be that $(\rho^*, t^*) = \left(\overline{\rho}, \overline{t}\right)$ which means that conditions $i)$ through $v)$ in the statement of proposition 27 hold.

Now consider the case where (ρ^*, t^*) does not solve the relaxed problem. That must be because (ρ^*, t^*) does not satisfy condition (A.8) and so, condition (A.8) must bind. Let (ρ^r, t^r) be a solution of the relaxed problem.

Observation 1: $t_n^r(v) = c$ for all n if $\rho^r(v) = 0$.

It is irrelevant what the transfer vector if the good is not provided, so I arbitrarily set each transfer equal to c.

Observation 2: $\sum_{n=1}^{N} t_n^r(v) = \widehat{c}$.

If not, it would be possible to decrease the transfers of $\overline{v}$ type agents. This would increase welfare and would satisfy conditions (A.8), (3.3) and (3.4).

Observation 3: $\overline{t}_n^r(v) = \overline{t}_{n'}^r(v)$ if $v_n = v_{n'}$.

If $\overline{t}_n^r(v) > \overline{t}_{n'}^r(v)$, decreasing the former and increasing the latter in such a way that the sum is the same would still satisfy all constraints while increasing expected welfare given that W is strictly concave. The increase in the expected welfare is strict if $\rho^r(v) > 0$.

Observation 4: $\rho^r(v) = 0$ if $i(v)\overline{v} + (N - i(v))\underline{v} < \widehat{c}$.

If not, some agent's ex post utility would be negative.

Observation 5: $\rho^r(\overline{v}, ..., \overline{v}) = 1$.

If not, it would be possible to increase $\rho^r(\overline{v},...,\overline{v})$ and still satisfy condition (A.8) while increasing the expected welfare.

Notice that observations 2 and 3 imply statements i) and ii) of the proposition as long $(\rho^r, t^r) = \left(\overline{\rho}, \overline{t}\right)$. Following observation 3 we denote $\overline{t}_n^r(v)$ by $\tau_{\overline{v}}^r(v)$ whenever $v_n = \overline{v}$ and by $\tau_{\underline{v}}^r(v)$ whenever $v_n = \underline{v}$. Then, following observation 2, we have that

$$i(v)\,\tau_{\overline{v}}^r(v) + (N - i(v))\,\tau_{\underline{v}}^r(v) = Nc \text{ if } \rho^r(v) > 0$$

which implies that

$$\tau_{\underline{v}}^r(v) = \frac{N}{(N - i(v))}c - \frac{i(v)}{(N - i(v))}\tau_{\overline{v}}^r(v) \text{ if } \rho^r(v) > 0$$

I am now able to write condition (A.8) simply as a function of ρ^r and of $\tau_{\overline{v}}^r$ as follows:

$$p^{N-1}N(\overline{v} - c) \geq \sum_{v \notin \{(\overline{v},...,\overline{v}),(\underline{v},...,\underline{v})\}} \rho^r(v)\,p^{i(v)-1}(1-p)^{N-i(v)-1}\left(pN(\overline{v} - c) - i(v)(\overline{v} - \tau_{\overline{v}}^r(v))\right) \tag{A.10}$$

Notice that the ex-post individual rationality constraints on $\overline{v}$ type agents do not bind and the $\underline{v}$ type agents' can be written as

$$\tau_{\overline{v}}^r(v) \geq \frac{N}{i(v)}c - \frac{(N - i(v))}{i(v)}\underline{v} \tag{A.11}$$

Fix any $v \notin \{(\overline{v},...,\overline{v}),(\underline{v},...,\underline{v})\}$. By analyzing the FOC associated with the choice of $\tau_{\overline{v}}^r(v)$ we get that, whenever $\rho^r(v) > 0$,

$$p(1-p)\left(\frac{\partial W}{\partial u_{\underline{v}}}-\frac{\partial W}{\partial u_{\overline{v}}}\right)-\mu+\frac{\eta(v)}{p^{i(v)-1}(1-p)^{N-i(v)-1}\rho^{r}(v)\,i(v)}=0 \tag{A.12}$$

where $\frac{\partial W}{\partial u_{\overline{v}}}\equiv\frac{\partial W}{\partial u_n}$ for $n:v_n=\overline{v}$, $\frac{\partial W}{\partial u_{\underline{v}}}\equiv\frac{\partial W}{\partial u_n}$ for $n:v_n=\underline{v}$, $\mu>0$ is the multiplier associated with (A.10) and $\eta(v)\geq 0$ is the multiplier associated with (A.11).

Notice that, given that W is strictly concave there is a unique $\widehat{\tau}(\mu)$ such that

$$\left(\frac{\partial W}{\partial u_{\underline{v}}}-\frac{\partial W}{\partial u_{\overline{v}}}\right)@\left(\left(\begin{array}{c} i(v)\otimes(\overline{v}-\widehat{\tau}(\mu))\oplus \\ (N-i(v))\otimes\left(\underline{v}-\frac{N}{(N-i(v))}c+\frac{i(v)}{(N-i(v))}\widehat{\tau}(\mu)\right)\end{array}\right)\right)=\frac{\mu}{p(1-p)}$$

where $W\left(i\otimes u_{\overline{v}}\oplus(N-i)\otimes u_{\underline{v}}\right)$ is interpreted as the welfare when there are i high valuation agents with utility $u_{\overline{v}}$ and $N-i$ low valuation agents with utility $u_{\underline{v}}$. Hence, if $\widehat{\tau}\geq\frac{N}{i(v)}c-\frac{(N-i(v))}{i(v)}\underline{v}$ then, $\tau_{\overline{v}}^{r}(v)=\widehat{\tau}$; otherwise $\tau_{\overline{v}}^{r}(v)=\frac{N}{i(v)}c-\frac{(N-i(v))}{i(v)}\underline{v}$. Either way, $\tau_{\overline{v}}^{r}(v)$ only depends on $i(v)$ as long as $\rho^{r}(v)>0$.

In an arbitrary optimal solution of the relaxed problem, it is possible that $\rho^{r}(v)>0$ and $\rho^{r}(v')=0$ while $i(v)=i(v')$. However, by shifting half of the weight in $\rho^{r}(v)$ to $\rho^{r}(v')$ (so that they both have the same weight) we keep the expected welfare constant and all constraints of the relaxed problem hold. Hence, we define as $\overline{\rho}$ the solution of the relaxed problem that is such that $\overline{\rho}(v)=\overline{\rho}(v')$ whenever $i(v)=i(v')$.

Given that $\tau_{\overline{v}}^{r}(v)$ only depends on $i(v)$ property iii) follows. Property iv) follows by noticing that $\widehat{\tau}(0)=t_n^{*}(v)$ for n such that $v_n=\overline{v}$.

Now consider the FOC associated with $\overline{\rho}$ we get

$$\left(\begin{array}{c} p(1-p)\,W\left(i(v)\otimes u_{\overline{v}}\oplus(N-i(v))\otimes u_{\underline{v}}\right) \\ -\mu\left(pN(\overline{v}-c)-i(v)(\overline{v}-\tau_{\overline{v}}^{r}(v))\right)+\frac{\overline{\xi}(v)-\underline{\xi}(v)}{p^{i(v)-1}(1-p)^{N-i(v)-1}}\end{array}\right)=0 \tag{A.13}$$

Given condition (A.13) $\overline{\rho}$ will have the form of a threshold rule. Given that $\tau_{\overline{v}}^{r}(v) \geq \tau_{\overline{v}}^{r}(v')$ for all v, v' such that $i(v) > i(v')$ and $W\left(i(v) \otimes u_{\overline{v}} \oplus (N - i(v)) \otimes u_{\underline{v}}\right)$ is increasing with $i(v)$, $\overline{\rho}$ is increasing. Notice that the threshold by $\overline{i} \geq \widehat{i}$ given that $\mu \geq 0$. These observations imply property v).

Finally, the only thing left is to show that $\left(\overline{\rho}, \overline{t}\right)$ is also the solution of the original problem. For that it is enough to show that it satisfies all incentive constraints that are given by condition (3.2). Notice that $A(\overline{\rho}) \geq 0$ which implies that condition (A.9) holds. Given that $\left(\overline{\rho}, \overline{t}\right)$ does not depend on the identity of the agents, the individual incentive constraints are also met which shows that $\left(\overline{\rho}, \overline{t}\right)$ is indeed a solution of the original problem.

A.3.2. Proof of Proposition 30

First, I show that it is enough to consider one "high" and one "low" message for any agent n, where "high" messages are messages that are sent only by high types and "low" messages are sent only by low types. Take an agent n and suppose he is sending two high messages. Each of the two messages imply the same posterior belief over all agents' types so we can simply treat those two messages as a single one. The same argument follows for low messages.

Now, I show that it is enough to consider a single "mixed" message per agent, where "mixed" messages are messages that are sent by both types - $\overline{v}$ and $\underline{v}$. Take any agent n and suppose he sends two mixed messages. If the two mixed messages induce the same posterior, then the previous argument follows. Suppose not and say that $p_n(m_n') > p_n(m_n'')$. Because both messages are mixed, the probability the public good is provided is the same under both messages. So, for agent n to be indifferent between the two, it must be that the expected transfer of both messages is exactly the same. This means that, as long as the public good gets provided, the transfer player n pays regardless of whether he sends messages m_n' or m_n'' is always $\underline{v}$. I now propose to join the two messages into one message $\widetilde{m}_n$. In particular, say

that $s_n^{m_n}(v_n)$ is the probability that player n with type v_n plays message m_n. Then have $s_n^{\widetilde{m}_n}(v_n) = s_n^{m'_n}(v_n) + s_n^{m''_n}(v_n)$ for $v_n \in \{\underline{v}, \overline{v}\}$. Notice that $p_n(\widetilde{m}_n) \in [p_n(m''_n), p_n(m'_n)]$. This means that all transfers associated with $\widetilde{m}_n$ are the same as with m'_n and with m''_n which means that such a switch to $\widetilde{m}_n$ leaves all agents the same and the expected welfare unchanged.

Finally, I show that we cannot have low types sending both a low message and a mixed message. The argument is similar to the previous one except that, by lemma 29)iv), the expected transfer associated with a low message is always strictly smaller than with a mixed message and so the low message strictly dominates the mixed one.

The last point is to make sure that an equilibrium always exists which is guaranteed by the fact that if $s = (1, ..., 1)$ then $\left(\widetilde{\sigma}(s), \widetilde{\xi}\right)$ is a BD equilibrium, where $\widetilde{\xi}$ is described in Lemma 29.

A.3.3. Proof of Proposition 32

Notice that the message space for each agent consists of a report about his type and a choice of a real number from the interval $[0,1]$. In this framework we can define, for each $m \in M$, $r(m) \in \{\overline{v}, \underline{v}\}^N$ to be the report part of message m and $z(m) \in [0,1]^N$ to be the part of message m that refers to the real number choice so that $m = (r(m), z(m))$.

Let $\left(\widehat{\sigma}, \widehat{\phi}\right)$ be such that:

$$\widehat{\sigma}_n^{v_n}(a,b) = \begin{cases} 1 \text{ if } a = v_n \text{ for all } b \in [0,1] \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } v_n \in \{\overline{v}, \underline{v}\} \text{ and for all } n$$

which means that agents report their type truthfully and choose all real numbers in $[0,1]$

with equal probability. Also let

$$\widehat{\phi}_n^{\overline{v},m}(a) = \begin{cases} 1 \text{ if } a = \bar{t}_n(r(m)) \text{ and } g(z(m)) \leq \overline{\rho}(r(m)) \\ 1 \text{ if } a = 0 \text{ and } g(z(m)) > \overline{\rho}(r(m)) \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } m \text{ and } n$$

and

$$\widehat{\phi}_n^{\underline{v},m}(a) = \begin{cases} 1 \text{ if } a = \bar{t}_n(r(m)),\ m_n = \underline{v} \text{ and } g(z(m)) \leq \overline{\rho}(r(m)) \\ 1 \text{ if } a = 0 \text{ and either } m_n = \overline{v} \text{ or } g(z(m)) > \overline{\rho}(r(m)) \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } m \text{ and } n$$

where $g : [0,1]^N \to [0,1]$ is an extension of the binary matching function in Matthews and Postlewaite (1989). The idea is that we are able to replicate $\overline{\rho}$ without resorting to mixing the contributions each agent makes as that would lead to the waste of resources, i.e. the sum of contributions would sometimes exceed $\widehat{c}$ or be in between 0 and $\widehat{c}$.

Now, I explain what the g function's role is. Let $f : [0,1]^2 \to [0,1]$ be defined as in Matthews and Postlewaite (1989): $f(z_1, z_2)$ is equal to the number whose binary expansion has a "1" in the nth position if and only if the binary expansion of z_1 has the same digit in nth position as does the binary expansion of z_2. The idea is that the distribution of f is uniform on $[0,1]$ as long as either z_1 or z_2 are also uniform on $[0,1]$ and are independent. We define g to be composite function of all these binary functions as follows:

Define $\widehat{g}^2 = f$ and $\widehat{g}^n : [0,1]^n \to [0,1]$ such that

$$\widehat{g}^n(z_1, .., z_{n-1}, z_n) = f\left(\widehat{g}^{n-1}(z_1, .., z_{n-1}), z_n\right)$$

for all $n \in \{3, ..., N\}$. Then $g \equiv \widehat{g}^N$. Notice that, if $z_n \sim U(0,1)$ for at least $N-1$ elements and all z_n are independent then g's distribution is also uniform on $(0,1)$.

To see that the profile $\left(\widehat{\sigma}, \widehat{\phi}\right)$ implements $\left(\overline{\rho}, \overline{t}\right)$, notice that $\widehat{\sigma}$ implies there is truthful

reporting among the agents. In order for the probability the public good is provided to match the one specified in $\overline{\rho}$ we have agents reporting a real number in $[0,1]$. By combining all the numbers reported through function g we are able to coordinate play and match $\overline{\rho}$. To match $\overline{t}$ we define the second period transfers accordingly.

Now I argue that the profile $\left(\widehat{\sigma},\widehat{\phi}\right)$ is indeed an anarchic equilibrium. Given any m, it follows that $\widehat{\phi}$ induces a Bayes-Nash equilibrium of the contribution game. As for the reporting stage, notice that if an agent with type $\overline{v}$ chooses to report they are of type $\underline{v}$, they would act as if they really were of type $\underline{v}$ in the second period. This means that, because $\left(\overline{\rho},\overline{t}\right)$ is an incentive compatible allocation, such misreport would not be strictly preferred. Agents with type $\underline{v}$ have an expected payoff of 0 if they misreport. Hence, they too choose not to misreport, given they would always select a transfer of 0 in the second period. Finally, the specific g function also guarantees that all agents are indifferent between selecting any number from $[0,1]$ as long as everyone else is selecting all numbers with equal probability.

A.3.4. Proof of Proposition 33

First, I show that, for any $k>1$ and for any message m, contributing $t_{\widehat{n}} \geq \frac{2}{k}c$ is strictly dominated by $t'_{\widehat{n}} = t_{\widehat{n}} - \frac{2}{k}c$ for any agent $\widehat{n}$, regardless of his type. Notice that, by reducing his transfer in $\frac{2}{k}c$, the expected units of the public good that are provided go down at most by $\frac{1}{k}$. This is because, for g units to be provided, it must be that $t_{\widehat{n}} + \sum_{n\neq\widehat{n}} t_n \geq g\widehat{c}$ which implies that $t_{\widehat{n}} + \sum_{n\neq\widehat{n}} t_n - \frac{2}{k}c \geq \left(g-\frac{1}{k}\right)\widehat{c}$ because $N \geq 2$. Hence, the difference in the expected utility of agent $\widehat{n}$ is given by $2c - v_n > 0$ for all $v_n \in \{\underline{v},\overline{v}\}$.

If a transfer of $t_{\widehat{n}} \geq \frac{2}{k}c$ is strictly dominated it means that there are only two types of messages: the ones after which $\frac{1}{k}$ units of the good are provided and the ones after which there is no provision of the public good. Hence, WLOG, it is possible to treat the problem of finding the highest expected welfare anarchic equilibrium as having only 2 possibilities for the public good - 0 or $\frac{1}{k}$. But that was the problem I have solved in the previous section

- see Propositions 27 and 32. The property of the solution that is of relevance for this proof is that, after any message m, all agents select a unique transfer, i.e. they do not randomize.

Fix any integer $\overrightarrow{k} > 1$ and profile $\left(\sigma, \overrightarrow{\phi}\right)$ such that $\left(\sigma, \overrightarrow{\phi}\right)$ is the highest expected welfare anarchic equilibrium when $k = \overrightarrow{k}$ and $\overrightarrow{\phi}$ is such that agents do not randomize over transfers in the second stage.

I show that, for all integers $\overleftarrow{k} < \overrightarrow{k}$, there is always a profile $\left(\sigma, \overleftarrow{\phi}\right)$ such that $\left(\sigma, \overleftarrow{\phi}\right)$ is an anarchic equilibrium when $k = \overleftarrow{k}$ and its outcome induces a higher expected welfare than $\left(\sigma, \overrightarrow{\phi}\right)$, which shows the result.

Define $\overleftarrow{\phi}$ as follows. Let $\overleftarrow{\phi}_n^{v_n,m}(a) = \overrightarrow{\phi}_n^{v_n,m}\left(\frac{\overrightarrow{k}}{\overleftarrow{k}}a\right)$ for all v_n, m and n - all transfers after any message are multiplied by a factor of $\frac{\overrightarrow{k}}{\overleftarrow{k}}$. Now I show that $\left(\sigma, \overleftarrow{\phi}\right)$ is an anarchic equilibrium.

First, I show that no agent wants to deviate on the second period by selecting a different transfer than the one specified in $\overleftarrow{\phi}$. Consider the payoff agent n receives by choosing transfer $\overleftarrow{t}_n$ after observing message m, given profile $\left(\sigma, \overleftarrow{\phi}\right)$:

$$\overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m, \overleftarrow{t}_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{t}_n\right)$$

where $\overleftarrow{\Pr}$ stands for the probability given the beliefs induced by $\left(\sigma, \overleftarrow{\phi}\right)$. Now take $\overrightarrow{t}_n$ such that $\overleftarrow{t}_n = \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{t}_n$. Notice that

$$\overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m, \overleftarrow{t}_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{t}_n\right) = \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m, \overrightarrow{t}_n\right\}\left(\frac{1}{\overrightarrow{k}}v_n - \overrightarrow{t}_n\right)$$

because

$$\overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m, \overleftarrow{t}_n\right\} = \overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m, \overrightarrow{t}_n\right\}$$

where $\overrightarrow{\Pr}$ stands for the probability given the beliefs induced by $\left(\sigma, \overrightarrow{\phi}\right)$.

Suppose agent n, after some message m, strictly prefers to deviate in the second stage and make a transfer $\overleftarrow{t}'_n$ such that $\overleftarrow{\phi}\left(\overleftarrow{t}'_n\right) = 0$. Then, it must be that

$$\begin{aligned}
\overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m, \overleftarrow{t}_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{t}_n\right) &< \overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m, \overleftarrow{t}'_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{t}'_n\right) \\
&= \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m, \overrightarrow{t}'_n\right\}\left(\frac{1}{\overrightarrow{k}}v_n - \overrightarrow{t}'_n\right) \\
&\leq \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m, \overrightarrow{t}_n\right\}\left(\frac{1}{\overrightarrow{k}}v_n - \overrightarrow{t}_n\right) \\
&= \overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m, \overleftarrow{t}_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{t}_n\right)
\end{aligned}$$

where $\overrightarrow{t}'_n$ is such that $\overleftarrow{t}'_n = \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{t}'_n$. Given that the previous derivation leads to a contradiction, no agent wants to deviate in the second period.

Finally, using a similar argument, I show that no agent wishes to misreport. Notice that by playing a given message m_n, agent n's expected utility is given by

$$\overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{E}_n\left(t_n|m_n\right)\right) = \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m_n\right\}\left(\frac{1}{\overrightarrow{k}}v_n - \overrightarrow{E}_n\left(t_n|m_n\right)\right)$$

where $\overleftarrow{E}_n$ and $\overrightarrow{E}_n$ stand for the expected value agent n forms, given the beliefs induced by $\left(\sigma, \overleftarrow{\phi}\right)$ and $\left(\sigma, \overrightarrow{\phi}\right)$ respectively. Suppose agent n deviates and strictly prefers to report m'_n such that $\sigma\left(m'_n\right) = 0$. Then it must be that

$$\begin{aligned}
\overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{E}_n\left(t_n|m_n\right)\right) &< \overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m'_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{E}_n\left(t_n|m'_n\right)\right) \\
&= \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m'_n\right\}\left(\frac{1}{\overrightarrow{k}}v_n - \overrightarrow{E}_n\left(t_n|m'_n\right)\right) \\
&\leq \frac{\overrightarrow{k}}{\overleftarrow{k}}\overrightarrow{\Pr}\left\{g = \frac{1}{\overrightarrow{k}}|m_n\right\}\left(\frac{1}{\overrightarrow{k}}v_n - \overrightarrow{E}_n\left(t_n|m_n\right)\right) \\
&= \overleftarrow{\Pr}\left\{g = \frac{1}{\overleftarrow{k}}|m_n\right\}\left(\frac{1}{\overleftarrow{k}}v_n - \overleftarrow{E}_n\left(t_n|m_n\right)\right)
\end{aligned}$$

which is a contradiction. Hence, $\left(\sigma, \overleftarrow{\phi}\right)$ is an anarchic equilibrium if $k = \overleftarrow{k}$.

Finally, given that $\frac{\overrightarrow{k}}{\overleftarrow{k}} > 1$ the utility each agent receives after any message sent is higher under $\left(\sigma, \overleftarrow{\phi}\right)$ which proves the result given that W is increasing.

A.3.5. Proof of Proposition 34

The only thing I need to show is that, for any beliefs, the BD's decision is independent of k. Recall that, from i) and ii) of Lemma 29, we have that, when $k = 1$, the BD provides 1 unit of the good after any message m such that $i(m) \geq \widehat{i}$ and 0 units of the public good after all other messages. Notice that, for the latter ones, the BD will still not provide the public good as he has no way of funding them because

$$i\overline{v} + (N - i)\,\underline{v} < \widehat{c} \Rightarrow \frac{1}{k}\left(i\overline{v} + (N - i)\,\underline{v}\right) < \frac{1}{k}\widehat{c}$$

for any k.

Now consider all messages m such that $i(m) \geq \widehat{i}$ and suppose that the BD chooses to provide $\frac{\widehat{k}}{k}$ units of the public good, where $k > 1$ and $\widehat{k} < k$. Let $\widehat{y}(m)$ denote the optimal transfer scheme chosen by the BD in that case. Notice that

$$\frac{\widehat{k}}{k}v_n - \widehat{y}_n(m) \leq \frac{k}{\widehat{k}}\left(\frac{\widehat{k}}{k}v_n - \widehat{y}_n(m)\right) = v_n - \frac{k}{\widehat{k}}\widehat{y}_n(m)$$

where the inequality follows from the fact that $\frac{\widehat{k}}{k}v_n - \widehat{y}_n(m) \geq 0$ due to the ex-post individual rationality constraints and must be strict for some agent n. Hence, by providing 1 unit of the public good and setting a transfer of $\frac{k}{\widehat{k}}\widehat{y}_n(m)$ on each agent n, we make all agents better off (and at least one strictly so) and are able to fully fund the provision the public good, given that $\sum_{n=1}^{N}\frac{k}{\widehat{k}}\widehat{y}_n(m) = \widehat{c}$. We then have a contradiction because W is strictly increasing. Therefore, the BD always makes the same decision, for given beliefs, independently of k.

A.3.6. Proof of Proposition 36

In proposition 30, I show that $\left(\widetilde{\sigma}\left(s\right),\widetilde{\xi}\right)$ is a BD equilibrium that maximizes the expected welfare among all the BD equilibria for some $s \in [0,1]^N$ where $s \neq (0,...0)$ (for otherwise $\left(\overline{\rho},\overline{t}\right) = (\rho^*, t^*)$). In this proof, I present a mediator equilibrium $\left(\widehat{\sigma},\widehat{\zeta},\widetilde{\xi}\right)$ that strictly improves upon that BD equilibrium.

First, I present an intermediate step, where I show that there is a mediator equilibrium $\left(\widehat{\sigma},\widetilde{\zeta},\widetilde{\xi}\right)$ with truthful reporting that induces the same expected welfare as the BD equilibrium $\left(\widetilde{\sigma}\left(s\right),\widetilde{\xi}\right)$.

Let $M_n = \{L, H\}$ and set

$$\widehat{\sigma}_n^{v_n}\left(a\right) = \begin{cases} 1 \text{ if } v_n = \overline{v} \text{ and } a = H \\ 1 \text{ if } v_n = \underline{v} \text{ and } a = L \\ 0 \text{ otherwise} \end{cases}$$

so that agents are fully revealing their type to the mediator. Then have, for all m,

$$\widetilde{\zeta}^m\left(a\right) = \prod_{n=1}^{N} \left(\begin{array}{c} 1\left\{a_n = H\right\} 1\left\{m_n = H\right\}\left(1 - s_n\right) \\ +1\left\{a_n = L\right\}\left(1\left\{m_n = H\right\} s_n + 1\left\{m_n = L\right\}\right) \end{array} \right)$$

Notice that $\left(\widehat{\sigma},\widetilde{\zeta},\widetilde{\xi}\right)$ is a mediator equilibrium, where the distribution of messages the BD receives is equal to the BD equilibrium $\left(\widetilde{\sigma}\left(s\right),\widetilde{\xi}\right)$. The difference now is that the mixing between the messages is done not by the agents but by the mediator.

Now, I show there is a mediator equilibrium $\left(\widehat{\sigma},\widehat{\zeta},\widetilde{\xi}\right)$ that strictly improves upon the

mediator equilibrium $\left(\widehat{\sigma},\widetilde{\zeta},\widetilde{\xi}\right)$. Let

$$\widehat{\zeta}^{(H,...,H)}(a)=\begin{cases}\widetilde{\zeta}^{(H,...,H)}(a)+\sum_{b\in\Gamma}\widetilde{\zeta}^{(H,...,H)}(b) \text{ if } a=(H,...,H)\\ 0 \text{ if } \sum_{n=1}^{N}1\{a_n=H\}<\widehat{i}\\ \widetilde{\zeta}^{(H,...,H)}(a) \text{ otherwise}\end{cases}$$

where

$$\Gamma=\left\{b\in M:\sum_{n=1}^{N}1\{b_n=H\}<\widehat{i}\right\}$$

and $\widehat{\zeta}^{m}=\widetilde{\zeta}^{m}$ for all $m\neq(H,...,H)$. I claim that $\left(\widehat{\sigma},\widehat{\zeta},\widetilde{\xi}\right)$ is a mediator equilibrium that induces a strictly higher expected welfare. To see that the profile $\left(\widehat{\sigma},\widehat{\zeta},\widetilde{\xi}\right)$ is a mediator equilibrium notice that the posterior beliefs the BD holds for any message m are the same as with $\left(\widehat{\sigma},\widetilde{\zeta},\widetilde{\xi}\right)$ whenever the BD decided to provide the public good. For messages where the public good did not get provided, the BD now believes that all agents are more likely to have a low type which means he still does not provide the good in $\left(\widehat{\sigma},\widehat{\zeta},\widetilde{\xi}\right)$. So the payoffs per message the BD receives are the same. Finally, the expected utility of reporting L did not change while the expected utility of reporting H increased only for agents with a high type (because $\overline{v}>c>\underline{v}$) which means that all agents prefer to reveal their type then to misreport. The equilibrium $\left(\widehat{\sigma},\widehat{\zeta},\widetilde{\xi}\right)$ has a higher expected welfare because whenever all agents have a high type there is an increase of $\sum_{b\in\Gamma}\widetilde{\zeta}^{(H,...,H)}(b)$ in the probability that the good is provided and the transfers are ex-post optimal, as opposed to the good not being provided.

Bibliography

- Agastya, Murali, Flavio Menezes, and Kunal Sengupta. "Cheap talk, efficiency and egalitarian cost sharing in joint projects." Games and Economic Behavior 60.1 (2007): 1-19.

- Aumann, Robert J., Michael Maschler, and Richard E. Stearns. "Repeated Games of Incomplete Information: An Approach to the Non-Zero-Sum Game". Report to the US Arms Control and Disarmament Agency, Contract S.T. 143, prepared by Mathematica Inc., Princeton, NJ (1968).

- d'Aspremont, Claude, and Louis-Andre Gérard-Varet. "Incentives and incomplete information." Journal of Public economics 11.1 (1979): 25-45.

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

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

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

- Barbieri, Stefano. "Communication and Early Contributions." Journal of Public Economic Theory 14.3 (2012): 391-421.

- Barbieri, Stefano, and David A. Malueg. "Threshold uncertainty in the private-information subscription game." Journal of Public Economics 94.11 (2010): 848-861.

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

- Bénabou, Roland, and Jean Tirole. "Self-confidence and personal motivation." The Quarterly Journal of Economics 117.3 (2002): 871-915.

- Benoît, Jean-Pierre, and Juan Dubra. "Apparent overconfidence." Econometrica 79.5 (2011): 1591-1625.

- Besley, Timothy, and Stephen Coate. "Centralized versus decentralized provision of local public goods: a political economy approach." Journal of public economics 87.12 (2003): 2611-2637.

- 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.

- Brocas, Isabelle, and Juan D. Carrillo. "Information acquisition and choice under

uncertainty." Journal of Economics & Management Strategy 18, no. 2 (2009): 423-455.

- Buehler, Roger, Dale Griffin, and Michael Ross. "Exploring the" planning fallacy": Why people underestimate their task completion times." Journal of personality and social psychology 67 (1994): 366-366.

- Camerer, Colin, and Dan Lovallo. "Overconfidence and excess entry: An experimental approach." The American Economic Review 89.1 (1999): 306-318.

- Cho, Myeonghwan. "Externality and information asymmetry in the production of local public good." International Journal of Economic Theory 9 (2013): 177-201.

- Clark, Jeremy, and Lana Friesen. "Overconfidence in Forecasts of Own Performance: An Experimental Study." The Economic Journal 119.534 (2009): 229-251.

- 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.

- 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.

- Dunning, David, Chip Heath, and Jerry M. Suls. "Flawed self-assessment implications for health, education, and the workplace." Psychological science in the public interest 5.3 (2004): 69-106.

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

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

- Glasgow, Kristan L., et al. "Parenting styles, adolescents' attributions, and educational outcomes in nine heterogeneous high schools." Child development 68.3 (1997): 507-529.

- Green, Jerry R. and Jean-Jacques Laffont. "Incentives in Public Decision Making". North-Holland, Amsterdam (1979).

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

- Groves, Theodore. "Incentives in teams." Econometrica (1973): 617-631.

- Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich. "Inside the judicial mind." Cornell L. Rev. 86 (2000): 777.

- Hattie, John, and Helen Timperley. "The power of feedback." Review of educational research 77.1 (2007): 81-112.
- Hayek, Friedrich August. "The use of knowledge in society." The American economic review (1945): 519-530.
- Kaplow, Louis, and Steven Shavell. "Optimal Law Enforcement with Self-Reporting of Behavior." Journal of Political Economy 102.3 (1994).
- Kessler, Anke S. "Communication in Federal Politics: Universalism, Policy Uniformity, and the Optimal Allocation of Fiscal Authority." Journal of Political Economy, 122.4 (2014): 766:805.
- Kim, Jeong-Yoo. "Secrecy and fairness in plea bargaining with multiple defendants." Journal of Economics 96.3 (2009): 263-276.
- Kim, Jeong-Yoo. "Credible plea bargaining." European Journal of Law and Economics 29.3 (2010): 279-293.
- Klassen, Robert M. "A Cross-Cultural Investigation of the Efficacy Beliefs of South Asian Immigrant and Anglo Canadian Nonimmigrant Early Adolescents." Journal of Educational Psychology 96.4 (2004): 731.
- Klibanoff, Peter, and Michel Poitevin. "A Theory of (De)centralization." (Unpublished paper, Northwestern University (2013).
- Kobayashi, Bruce H. "Deterrence with multiple defendants: an explanation for" Unfair" plea bargains." The RAND Journal of Economics (1992): 507-517.
- Köszegi, Botond. "Ego utility, overconfidence, and task choice." Journal of the European Economic Association 4.4 (2006): 673-707.
- Laffont, Jean-Jacques, and Eric Maskin. "On the difficulty of attaining distributional goals with imperfect information about consumers." The Scandinavian Journal of Economics (1979): 227-237.
- Laudan, Larry. "Truth, error, and criminal law: an essay in legal epistemology." Cambridge University Press (2006).
- Lewis, Tracy R., and David EM Sappington. "Motivating wealth-constrained actors." American Economic Review (2000): 944-960.
- Lockwood, Ben. "Distributive politics and the costs of centralization." The Review of Economic Studies 69.2 (2002): 313-337.
- Mailath, George J., and Andrew Postlewaite. "Asymmetric information bargaining problems with many agents." The Review of Economic Studies 57.3 (1990): 351-367.

- Makris, Miltiadis. "Private provision of discrete public goods." Games and Economic Behavior 67.1 (2009): 292-299.
- Malmendier, Ulrike, and Geoffrey Tate. "CEO overconfidence and corporate investment." The journal of finance 60.6 (2005): 2661-2700.
- 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.
- Matthews, Steven A., and Andrew Postlewaite. "Pre-play communication in two-person sealed-bid double auctions." Journal of Economic Theory 48.1 (1989): 238-263.
- Meyer, John P., and Ian R. Gellatly. "Perceived performance norm as a mediator in the effect of assigned goal on personal goal and task performance." Journal of Applied Psychology 73.3 (1988): 410.
- Midjord, Rune. "Competitive Pressure and Job Interview Lying: A Game Theoretical Analysis", *mimeo* (2013).
- Moore, Don A., and Paul J. Healy. "The trouble with overconfidence." Psychological review 115.2 (2008): 502.
- Myerson, Roger B. "Incentive compatibility and the bargaining problem." Econometrica: journal of the Econometric Society (1979): 61-73.
- Myerson, Roger B. "Mechanism design by an informed principal." Econometrica: Journal of the Econometric Society (1983): 1767-1797.
- Mylovanov, Tymofiy, and Andriy Zapechelnyuk. "Mechanism Design with ex-post Verification and Limited Punishments", *mimeo* (2014).
- Oates, Wallace E. "Fiscal federalism." (1972). Harcort, Brace, Jovanovich, New York.
- Posner, Richard A. "An economic approach to the law of evidence." Stanford Law Review (1999): 1477-1546.
- Samuelson, Paul A. "The pure theory of public expenditure." The review of economics and statistics (1954): 387-389.
- Santobello v. New York, 404 U.S. 257 (1971), 261.
- Schmidt, Klaus M. "The costs and benefits of privatization: an incomplete contracts approach." Journal of Law, Economics, and Organization 12.1 (1996): 1-24.
- Scott, Robert E., and William J. Stuntz. "Plea bargaining as contract." Yale Law Journal (1992): 1909-1968.

- Siegel, Ron and Bruno Strulovici. "On the design of Criminal Trials: The Benefits of a Three-Verdict System", *mimeo* (2015).

- Smith, Herbert L., and Brian Powell. "Great expectations: Variations in income expectations among college seniors." Sociology of Education (1990): 194-207.

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

- Svenson, Ola. "Are we all less risky and more skillful than our fellow drivers?." Acta Psychologica 47.2 (1981): 143-148.

- Taylor, Shelley E., and Jonathon D. Brown. "Illusion and well-being: a social psychological perspective on mental health." Psychological bulletin 103.2 (1988): 193.

- 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.

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

- Benabou, Roland, and Jean Tirole. "Intrinsic and extrinsic motivation." The Review of Economic Studies 70.3 (2003): 489-520.

- Zabojnik, Jan. "A model of rational bias in self-assessments." Economic Theory 23.2 (2004): 259-282.