Technische Universität München

TUM School of Management

Lehrstuhl für Volkswirtschaftslehre - Finanzwissenschaft und Industrieökonomik

# The Impact of Other-Regarding Preferences on Moral Hazard and Adverse Selection

Thomas Daske

Vollständiger Abdruck der von der Fakultät für Wirtschaftswissenschaften der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Volkswirtschaftslehre (Dr. oec. publ.) genehmigten Dissertation.

| | |
|---|---|
| Vorsitzender: | Prof. Dr. Martin Moog |
| Prüfer der Dissertation: | 1. Prof. Dr. Robert K. Frhr. von Weizsäcker |
| | 2. Prof. Dr. Michael Kurschilgen |

Die Dissertation wurde am 27.06.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Wirtschaftswissenschaften am 15.11.2017 angenommen.

## Abstract

This dissertation explores the role of other-regarding preferences in the form of altruism, spite, or status considerations in the resolution of adverse selection and moral hazard. Two chapters reflect upon the implications of asymmetric information on other-regarding preferences for incentive mechanism design in general and human resource management in specific. A third chapter reflects upon how moral hazard in the presence of other-regarding preferences may generate and shape peoples ethical convictions regarding their abidance by social norms and formal law. Implications for public economic policy are drawn.



## Zusammenfassung

Diese Dissertation untersucht die Rolle sozialer Präferenzen in der Form von Gunst, Missgunst oder Statusstreben für die Aufhebung von adverser Selektion und moralischem Risiko (moral hazard). Zwei Kapitel diskutieren die Konsequenzen von asymmetrischer Information ber soziale Präferenzen fr das Design von Anreizmechanismen im Allgemeinen und für Problemstellungen des Personalmanagements im Besonderen. Ein drittes Kapitel diskutiert, wie moralisches Risiko in Anwesenheit sozialer Präferenzen zur (Aus-)Prägung ethischer Überzeugungen im Hinblick auf die Einhaltung sozialer Normen führen kann. Politikimplikationen werden jeweils aufgezeigt.

## Acknowledgements

Many people contributed to the progress of my thesis in one way or another.

First and foremost, I would like to thank my thesis supervisor Robert von Weizsäcker. I have benefited greatly from his advice, encouragement, and support as well as from inspiring conversations about and beyond economics. I am grateful to Michael Kurschilgen for readily agreeing to be my second supervisor as well as for very helpful conversations. For his willingness to be part of my dissertation committee, I thank Martin Moog.

At the Chair of Economics at TUM School of Management, I have benefited from a supportive and creative team, for which I thank my colleagues Melanie Bader, Adrian Berger, Christoph Gschnaidtner, Julian Hackinger, Florian Inderst, Leonard Przybilla, Martine Pütz, and Alexander Schlimm. I am particularly grateful to my colleagues Christian Feilcke and Christoph March for inspiring debates, their ever-helpful advice, and their encouragement. I am indebted to Kai Konrad for his advice, fruitful discussions, and for giving me the opportunity to benefit from the research environment at the Max Planck Institute for Tax Law and Public Finance. For inspiring discussions and their support, I would like to thank Sabine Aresin, Anne-Kathrin Bronsert, Elias Brumm, Nadja Dwenger, Mauritz von Einem, May Elsayyad, Aart Gerritsen, Luisa Herbst, Michael Hilmer, Erik Hornung, David Houser, Changxia Ke, Harald Lang, Philipp Meyer-Brauns, Rhea Molato, Florian Morath, Salmai Qari, Tim Stolper, and Fangfang Tan. I thank Klaus Schmidt for fruitful conversations and helpful advice. I also benefited greatly from discussing part of my work in his Micro Workshop at LMU Munich. I thank Aart Gerritsen and Vai-Lam Mui for allowing me to make our joint work a part of my thesis.

Furthermore, I am grateful for helpful comments by participants of the 2016 Meeting of the European Public Choice Society in Freiburg, the Asian Meeting of the Econometric Society in Hong Kong, the European Meeting on Game Theory (SING13) in Paris, the Annual Congress of the Association for Public Economic Theory in Paris, and the Annual Congress of the International Institute of Public Finance in Tokyo.

Finally, I owe special thanks to my friends and family for their encouragement and their bravely listening to my research ideas and progress reports.

To my father

# Contents

# Introduction and Contribution

The scientific field broadly referred to as 'information economics' has deepened our understanding of the economic challenges associated with asymmetric information. Until recently, it has focused on information asymmetries with respect to *tangible* entities: Examples are the efficient provision of public goods if agents' valuations of those goods are private information (Groves and Ledyard, 1977); market failure in insurance markets if agents are privately informed about their 'risks' (Rothschild and Stiglitz, 1976); optimal taxation and, independently, the organization of labor markets if agents are privately informed about their 'skill' (Mirrlees, 1971, and Spence, 1973); and the feasibility of efficient trade if traders are privately informed about their valuations of the goods to be traded (Akerlof, 1970, and Myerson and Satterthwaite, 1983).

In reality, however, many of these settings might involve an additional dimension of information asymmetry, one that relates to *intangible* externalities associated with agents' other-regarding preferences. In many economic environments, people do not only care about their own material well-being but also about the material well-being of the people around them. In families or friendship networks, they might be intrinsically motivated to share (e.g., Becker 1976, 1981). At the workplace, or in society as a whole, they might rather care for their own advancement as compared to that of others (e.g., Easterlin, 1974, and Frank, 1985). Their other-regarding preferences, whether intrinsic or instrumental, will affect how people respond to incentives. They must be accounted for in order to render economic policies as well as the design of contracts and institutions efficient.

During the past decades, economists have spent increasing effort in aligning economic theory with individuals' other-regarding preferences, whether with regard to optimal taxation (starting with Boskin and Sheshinski, 1978) or in the range of human resource management (starting with Frank, 1984). Until recently, however, these strands of literature make an implicit assumption that seems questionable in most of the economic environments under investigation: namely, that those preferences are publicly observable and can thus be considered common knowledge among the relevant 'players'.

This thesis explores the role of other-regarding preferences in the resolution of adverse selection and moral hazard. Its main focus is put on the implications of asymmetric information on other-regarding preferences for the design of efficient incentive mechanisms: Chapter I, in a general mechanism design framework, considers other-regarding preferences that are private information to agents; Chapter II, in a more specific setup relevant to human resource management, considers other-regarding preferences that are common knowledge among agents (here, coworkers) but unobservable to the mechanism designer (here, a rent seeking principal). Chapter III takes a different view on other-regarding preferences and explores their role in the range of law and economics; it considers a scenario in which other-regarding preferences even induce a moral hazard problem between agents and investigates how agents' incentives to fight off this moral hazard may generate and shape their ethical convictions regarding the abidance by social norms and formal laws.

Specifically, Chapter I takes a general perspective on strategic interaction between individuals who privately assess the externalities their opponents might impose on them. These externalities can be associated with other-regarding preferences, but the model also extends to externalities that are tangible. Efficient mechanism design is explored under the assumption that agents' externality assessments and private payoffs, exclusive of externalities, are all subject to asymmetric information. Under reasonable assumptions, the following result is established: Let the allocation rule $f$ be the maximizer of a social welfare measure $W$ which satisfies the Pareto property; then $f$ is Bayesian implementable with an ex post budget-balanced mechanism *if and only if* $W$ sums private payoffs exclusive of externalities. By contrast, (nearly) any welfare judgment could be Bayesian implemented if one waived the requirement of budget balance or if agents' externality assessments were common knowledge. The result emphasizes the critical role of the welfare judgment inherent to the allocation rule if externality assessments are private information. Bayesian implementation of a welfare judgment inconsistent with externality-ignoring utilitarianism violates budget balance and, thus, involves incentive costs.

As an immediate application, Chapter I contrasts this result with the classical literature on 'cooperative' bargaining with: Even when allowing for side-payments, the renowned bargaining solutions proposed by Nash (1950) and Kalai (1977) cannot be Bayesian implemented if there is asymmetric information on bargainers' externality assessments.

Chapter I bridges three strands of literature: mechanism design in the presence of externalities, 'robust' mechanism design, and the measurement of social welfare. A growing literature, starting with Jehiel and Moldovanu (2001), investigates efficient implementation if there is asymmetric information on externalities. Typically, this literature takes a utilitarian point of view, requiring the allocation rule that is to be implemented to maximize the sum of agents payoffs *inclusive* of externalities. Chapter I adds to this literature a characterization of all those Pareto-efficient allocation rules that can be Bayesian implemented in a budget-balanced way. Under reasonable assumptions, there exists a unique such allocation rule, and it maximizes the sum of agents payoffs *exclusive* of externalities; I call the welfare judgment inherent to this allocation rule *externality-ignoring utilitarianism*.

The literature on 'robust' mechanism design accounts for Wilson's (1987) critique that game theory would rely too heavily on unrealistic common knowledge assumptions. Jehiel et al. (2006) have provided a negative result by showing that equilibrium concepts that require less common knowledge than Bayesian implementation can in almost all cases not be applied if there is asymmetric information on externalities. Bierbrauer and Netzer (2016) have shown that the 'Wilson doctrine' can at least be satisfied with regard to asymmetric information about agents' externality assessments, next to asymmetric information about agents' private payoffs: They consider a model in which externalities are associated with intention-based social preferences and are private information. Based on the renowned AGV-mechanism (due to Arrow, 1979, and d'Aspremont and Gérard-Varet, 1979), they provide sufficient conditions for externality-robust Bayesian implementation, meaning that agents and the mechanism designer do not need to have any knowledge of the distribution of social types. Chapter I contributes to this strand *necessary and sufficient* conditions for (efficient) externality-robust implementation. Indeed, the externality-robust AGV-mechanism is the unique mechanism that is both budget-balanced and Bayesian incentive compatible. That is, efficient implementation in the presence of asymmetric information on externality assessments even requires an externality-robust mechanism.

Finally, perhaps most importantly, Chapter I establishes a novel link between normative and positive theory. Unless the welfare judgment inherent to an allocation rule is consistent with externality-ignoring utilitarianism, its implementation violates budget balance and, thus, involves incentive costs. These incentive costs of welfare judgments can be interpreted as an incentive-compatibility constraint to the mechanism designer: Implementation of any other welfare judgment requires her to either subsidize agents or to accept that agents might incur losses. The model assumptions are satisfied by the CES-welfare measures proposed by Arrow (1973), which capture externality-sensitive utilitarianism as well as smooth approximations of 'Rawlsian justice' (Rawls, 1971), and by (smooth) social welfare measures that entail redistributive motives beyond utilitarianism. In this respect, the result is bad news for the proponents of public economic policies which result from non-utilitarian social welfare measures (e.g., Saez and Stantcheva, 2016).

A recent economic debate is concerned with the question of how workplace autonomy of employees would enhance a firm's productivity: Should employers abstain from (too much) control in order to not crowd out their employees' intrinsic motivation to conform with the firm's interests? While the literature tends to affirm this view (e.g., Falk and Kosfeld, 2006, Charness et al., 2012, and Flores-Fillol, Iranzo, and Mane, 2017), Chapter II puts it into perspective.

In many firms, production requires the division of staff into teams which, then, engage in parallel production. If only team performance is observable, moral hazard in teams is inevitable. This variant of moral hazard can be overcome or exacerbated by the interpersonal relationships among team members. I explore how the division of staff into teams should account for the agents' social network of interpersonal relationships. The main result states that the (potentially) unanimous preferences of staff for team composition can collide with efficient production. A universal mechanism guaranteeing efficiency while delegating responsibility for team assignment to the agents does not exist. Therefore, successful staffing requires knowledge of the interpersonal relationships at work and, at times, control instead of delegation.

Chapter II contributes to the field of personnel economics by linking the theory of moral hazard in teams to the theory of social and economic networks. Inspired and challenged by the seminal work of Holmstrom (1982), who showed that, whatever a sharing rule may look like, the free-rider problem in autonomous workgroups cannot be overcome, literally hundreds of studies have suggested explanations for the pervasive real-world phenomenon of autonomous workgroups. Chapter II adds a social network perspective to this literature. While most of the literature is concerned with rationales in support of efficient autonomous teamwork (e.g., Itoh, 1991, and Rotemberg, 1994), the key finding of Chapter II emphasizes the potential inefficiency of autonomous teamwork if teams form endogenously within a social network of other-regarding agents.

The theoretical literature on social and economic networks has mostly focused on the endogenous formation and stability of networks (e.g., Dutta, Ghosal, and Ray, 2005, and Page, Wooders, and Kamat, 2005), on networks of endogenous externalities among agents (e.g., Bramoullé and Kranton, 2007), and on equilibrium behavior in general 'network games' (e.g., Jackson and Wolinsky, 1996, and Galeotti et al., 2010). Chapter II adds a moral hazard perspective to social network theory.

Chapter III, which is joint work with Aart Gerritsen and Vai-Lam Mui, contributes to the growing economic literature that is dedicated to the role of parents in the transmission and evolution of social norms and ethics (e.g., Bisin and Verdier, 2001, Lindbeck and Nyberg, 2006, Tabellini, 2008, and Doepke and Zilibotti, 2017).

Law-abiding behavior is widely considered to depend not only on formal law enforcement but also on social norms as well as people's ethical convictions regarding law abidance. We investigate one plausible determinant of people's ethics of law abidance by considering the economic incentives of parents to bring up their children as law-abiding citizens. Altruistic parents might expect themselves to financially support their *grown up* children when those are convicted of illegal activities. Children, anticipating their parents' partial insurance provision, might therefore engage too much in illegal activities. Parental altruism thus breeds moral hazard. This variant of moral hazard provides parents with an incentive to instill an ethic of law abidance in their children while those are *adolescent*. We show that ethics formation is the result of a complex interplay of parents' own ethical convictions regarding law abidance, the extent of parental altruism, parents' assessments of their children's *legal* income prospects (as compared to their own wealth), as well as the determinants of formal law enforcement.

Specifically, we identify a non-monotonic relationship between expected intergenerational social mobility and parents' incentives to bring up their children as law-abiding citizens. Under intergenerational *downward* mobility, incentives are relatively *weak*, as parental support would have to be provided regardless of whether children succeeded in their illegal activities or were convicted and suffered from hefty fines. Ex ante, parents thus benefit from their children's noncompliance through a reduction in financial support. By contrast, under intergenerational *stagnation*, incentives are relatively *strong*, as parental support would be provided only in case of conviction, imposing the threat of moral hazard on parents. Under intergenerational *upward* mobility, incentives are *moderate*, as parents would not support their children even if those were convicted.

Our central question is how ethics formation will be affected by changes in law enforcement policy. We find that, under intergenerational stagnation as well as weak intergenerational upward or downward mobility, higher detection rates substitute for and thus crowd out ethics formation, whereas the effect of tougher punishment is ambiguous. On the other hand, under strong intergenerational upward or downward mobility, ethics formation is invariant to changes in formal law enforcement.

Our study links family economics to the literature on law and economics. Family economics (pioneered by Becker, 1974, 1976) investigates, among other things, how the conflicting preferences of family members affect the efficiency of resource allocation within families. In our model, parents face a particular variant of the 'Samaritan's dilemma' (Bruce and Waldman, 1990): Children might take advantage of their parents' partial insurance provision against conviction and punishment, leading to inefficiently high levels of illegal activity (from the family perspective). However, the family economics literature has not studied the incentives of parents to fight off this sort of moral hazard by instilling an ethic of law abidance in their children. We add this strategic dimension to it.

Following the pioneering work of Becker (1968) and Becker and Stigler (1974), there has emerged a large literature on the design of efficient formal law enforcement. Part of this literature reflects upon the interplay between formal law enforcement on the one hand and social norms or ethics of law abidance on the other, and is broadly referred to as 'crowding theory' (e.g., Frey and Jegen, 2001, Luttmer and Singhal, 2014, and Acemoglu and Jackson, 2017). We add to this literature a novel perspective: the role of intergenerational social mobility.

Policy implications of the above outlined findings are summarized in the final chapter of this thesis.

# Chapter I.

# Externality Assessments, Welfare Judgments, and Mechanism Design

## I.1. Introduction

The theory of mechanism design is devoted to the question of how to render collective action efficient if the agents involved hold private information—typically about their valuations of tangible assets. In many economic environments, however, this challenge is exacerbated by the fact that agents do also hold private information about their (rational or *ex post* irrational) assessments of the externalities that others might impose on them. These externalities can be tangible, for instance due to spillover effects between firms or local economies, or intangible—if agents derive (dis-)utility directly from how tangible assets are *distributed* among them.[1]

This study explores ex post Pareto-efficient (and, thus, ex post budget-balanced), mechanism design for two agents whose externality assessments and private payoffs are all subject to asymmetric information. Each agent's utility is taken as a weighted sum of her own payoff and her opponent's payoff, while the real-valued weight on the latter determines an agent's externality assessment, her *externality type*. An agent's payoff is additively separable in a numeraire good (money) and a payoff component (subject to the economic

[1]Agents might also derive (dis-)utility from—or change their preferences according to—the *process* through which final allocations are realized; see, e.g., Bowles and Hwang (2008). This line of reasoning is beyond the scope of the present study. Here, I take intangible externalities as outcome-dependent, being determined by agents' judgments about the final distribution of wealth.

environment under investigation) which is taken affine in her real-valued *payoff type*. An agent's externality type and payoff type are exogenously given, not perfectly correlated, and private information; types are independent across agents.—The central question is to what extent collective action can, or must, condition on agents' externality assessments in order to be ex post Pareto-efficient and incentivize agents to reveal their preferences truthfully.

With externalities taken tangible, the model captures bargaining between competing nations about scarce resources, with each nation having its private expectations about the benefit from that resource but also having its private expectations about the threat of the resource when being in the other nation's hands. Another example are neighboring municipalities negotiating harmonized public expenditure if there are spillovers from locally provided public goods.[2]

With externalities taken intangible, the model captures other-regarding preferences in the form of altruism, spite, or status. Altruism and spite are often deployed in the range of family economics.[3] The model captures bargaining problems like inheritance disputes and divorce battles, given that family members are privately informed about their valuations of the goods at stake (their payoff types) and about the extent to which they have come to despise each other (their externality types).[4] On the other hand, empirical studies have found that many, if not all, people care about their relative standing in society.[5] The model applies, for instance, to bargaining situations the outcomes of which will affect the income opportunities of bargainers, provided that the respective income expectations (payoff types) as well as relative standing considerations (externality types) are private information.

In order to implement ex post Pareto-efficient allocations, a mechanism provides agents with incentives such that they truthfully reveal their preferences in equilibrium.—What is

[2]This scenario has been analyzed by Harstad (2007), under the assumption of commonly known externalities though.

[3]E.g., Becker (1981).

[4]With regard to cross-ownership as outlined above, one can also think of two rulers in the cameralist era of European history who are related by marriage and negotiate the division of land.

[5]For empirical evidence on status considerations see, e.g., Clark, Frijters, and Shields (2008), Heffetz and Frank (2008), Tran and Zeckhauser (2012), and the survey by Weiss and Fershtman (1998). For a theoretical foundation of status preferences see, e.g., Bisin and Verdier (1998).

the appropriate equilibrium concept if there is asymmetric information about externality as well as payoff types?—This question is central not only to the design but also to the applicability of mechanisms, since different equilibrium concepts differ in their *common knowledge* assumptions about agents' information, preferences, and rationality. The aim to successively weaken common knowledge assumptions in game theory is sometimes referred to as the 'Wilson doctrine':

> "Game theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent it assumes other features to be common knowledge, such as one player's probability assessment about another's preferences or information.
>
> I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality." (Wilson, 1987)

The equilibrium concept with the weakest information requirement is that of dominant strategy implementation in the manner of Vickrey (1961), Clarke (1971), and Groves (1973). Unfortunately, with externalities, whether private information or common knowledge, dominant strategy implementation is typically not feasible. A weaker notion is that of ex post implementation, which requires that truthful revelation is each agent's best strategy in response to each and every realization of her opponents' (truthfully revealed) types. Under ex post implementation, knowledge of type distributions is not required. However, even if externality types are common knowledge, the imposition of budget balance restricts its applicability immensely.[6] The equilibrium concept I deploy is that of Bayesian implementation, which requires that truthful revelation maximizes each agent's von Neumann-Morgenstern (interim) expected utility provided all other agents reveal their types truthfully.[7] As Bayesian implementation collides with the 'Wilson doctrine', I will

[6] Bergemann and Morris (2005) show that Bayesian implementable allocation rules can, in many cases, no longer be ex post implemented when requiring budget balance.

[7] To be sure, the term *type* refers to the *pair* of an agent's externality and payoff type. Notice that a property which is possessed by the class of Bayesian implementable allocation rules is necessarily possessed by allocation rules that are ex post implementable.

put emphasis on how the assumption of common knowledge about the distribution of externality types can (and even must) be avoided.

In the environment under investigation, a mechanism specifies an *allocation rule*, specifying collective action based on the agents' preferences, and a *transfer scheme*, incentivizing agents to reveal those preferences. The challenge involved with private information about externality assessments is the following: Suppose the allocation rule conditions on externality assessments. Then the transfer scheme must elicit payoff types as well as externality types. However, through their externality assessments, agents internalize the distributive effects of the transfer scheme itself. Hence, the mechanism itself might deliver incentives to misrepresent preferences. Bayesian incentive compatibility demands counterbalance of these adverse incentives. Requiring budget balance further restricts the domain of adequate transfer schemes.

I show that the social welfare judgment inherent to an allocation rule is decisive for *whether* and *how* that allocation rule can be Bayesian implemented with a budget-balanced mechanism. Specifically, I obtain the following results.

By Proposition I.2, the renowned 'expected externality mechanism' (AGV-mechanism), due to Arrow (1979) and d'Aspremont and Gérard-Varet (1979), Bayesian implements in a budget-balanced way the allocation rule that, for each realization of types, maximizes the sum of private payoffs exclusive of externalities. These allocations are Pareto-efficient if each agent's marginal utility from her own payoff exceeds her marginal (dis-)utility from her opponent's payoff. The AGV-mechanism is *externality-robust* in the sense that it requires neither agents nor the mechanism designer to have any knowledge of the statistical distribution of externality types.

I then ask for conditions that an ex post Pareto-efficient allocation rule *must* satisfy in order to be Bayesian implementable with a budget-balanced mechanism. For this purpose, I introduce the notions of *sensitive allocation rules* and *strong Bayesian implementability.*

An allocation rule will be called *sensitive* if, in the respective economic environment, it is the unique maximizer of a social welfare measure which satisfies the Pareto property. Furthermore, a sensitive allocation rule is required to be *non-constant* in payoff types and to be *symmetric* in the sense that the effect of an increase in one agent's externality or payoff type on the other agent's private payoff is qualitatively similar for both agents. *Non-constancy* reflects strong, or 'sensitive', social welfare judgments of the mechanism designer, as it implies that she is not indifferent to even small changes in payoff types.[8]

An allocation rule will be called *strongly* Bayesian implementable if, for *any* set of (non-degenerate) type distributions, there exists a mechanism that Bayesian implements it. That is, strongly Bayesian implementable allocation rules may not condition on the specifics of type distributions. This requirement accounts for the 'Wilson doctrine' in so far as it avoids making common knowledge assumptions from the outset. By Proposition I.2, the allocation rule associated with externality-ignoring utilitarianism is sensitive and strongly Bayesian implementable.

I show that the converse of Proposition I.2 is also true if one asks for strong Bayesian implementation of sensitive allocation rules, which yields the following equivalence (Theorem I.1): A sensitive allocation rule can be strongly Bayesian implemented with a budget-balanced mechanism *if and only if* it maximizes the sum of private payoffs exclusive of externalities; I call the social welfare judgment inherent to these allocations *externality-ignoring utilitarianism.* The respective mechanism takes the form of the AGV-mechanism.

Loosely speaking, a sensitive allocation rule can be strongly Bayesian implemented in a budget-balanced way if and only if it results from a form of utilitarianism that approves individual achievements but ignores 'help' or 'harm' from others. Implementation of a social welfare judgment inconsistent with externality-ignoring utilitarianism violates budget balance and thus requires either an external source of money or that 'money is burned'. The associated costs can be interpreted as the incentive costs of the social welfare judgment. Furthermore, costless implementation of a sensitive allocation rule requires an externality-robust mechanism; all mechanisms having this property are of AGV-type. That is, the requirement of externality robustness does not only serve the purpose of satisfying the 'Wilson doctrine' but is even necessary from a welfarist point of view.

Finally, I outline the antagonistic roles of social welfare judgments and budget balance. Theorem I.2 shows that, even with asymmetric information about externality assessments,

[8] Examples of sensitive social welfare measures are given by utilitarian welfare, either *inclusive* or *exclusive* of externalities. When restricting the economic environment to linear utilities and non-negative externalities, several classical social welfare measures qualify as sensitive; they are listed in Proposition I.1.

nearly any social welfare judgment can be Bayesian implemented if one waives the requirement of budget balance. On the other hand, with privately observed payoff types but common knowledge of externality types, nearly any allocation rule can be Bayesian implementable in a budget-balanced way (Theorem I.3). Hence, it is not externality assessments per se that render social welfare judgments critical but rather the asymmetry of information about them combined with the efficiency request of budget balance.

The chapter proceeds as follows. Section I.2 reviews the related literature. Section I.3 outlines the basic model. Section I.4 identifies conditions that are necessary and sufficient for ex post Pareto-efficient Bayesian implementation; the central result on the allocative implications of social welfare judgments is obtained. Section I.5 interprets results for strategic bargaining under incomplete information. Section I.6 expands the central result to social welfare measures that incorporate the redistributive effects of the transfer scheme itself. Section I.7 concludes.

## I.2. Related Literature

This chapter relates to three strands of literature: 'robust' implementation, implementation in the presence of externalities, and the measurement of social welfare.

In order to come by the criticism pointed at unrealistic common knowledge assumptions (Wilson, 1987), many studies have characterized conditions under which Bayesian implementable allocation rules are ex post or even dominant strategy implementable.[9] Jehiel et al. (2006) consider a model framework that entails the one presented here, with the exception that agents do not internalize the distributive effects of transfers. They show that only those allocation rules can be ex post implemented that appoint the very same allocation for any realization of types. The implications of their result for the questions addressed here are discussed in detail at the end of Section I.4.

[9] E.g., Mookherjee and Reichelstein (1992), Dasgupta and Maskin (2000), Bergemann and Morris (2005, 2011), Chung and Ely (2007), Gershkov et al. (2013).

Several studies have explored ex post or Bayesian implementation under the assumption that externality assessments are common knowledge.[10] The present study considers Bayesian implementation while relaxing this assumption.[11] The studies closest to the present one are those of Jehiel and Moldovanu (2001) and Bierbrauer and Netzer (2016).

Jehiel and Moldovanu (2001) investigate the feasibility of 'efficient' Bayesian implementation in the presence of (allocative or informative) externalities.[12] In their model, each agent $i$ is privately informed about her private payoff, exclusive of externalities, and about the externality she imposes on another agent $j$. Agent $j$'s externality type, in the language of the present study, is assumed common knowledge. The present study expands the work of Jehiel and Moldovanu (2001) to the extent that it takes the externality of $i$ on $j$ as a composite of two pieces of private information, one held by $i$, the other one held by $j$. However, in order to expose the critical role of welfare judgments, attention is restricted to more specific economic environments.

Bierbrauer and Netzer (2016) explore the design of mechanisms for agents who exhibit intention-based social preferences in the manner of Rabin (1993). In a novel attempt, they allow for private information on social types and identify *sufficient* conditions for externality-robust Bayesian implementation.[13] The present study, in a slightly different setting, supplements their work by asking for *necessary and sufficient* conditions for budget-balanced Bayesian implementation.

This study bridges normative and positive theory based on incentive theoretical grounds. With regard to 'efficient' implementation, the mechanism design literature typically takes a utilitarian view. In the presence of externalities, the allocation rule is typically taken to maximize the sum of private payoffs *inclusive* of externalities (e.g., Jehiel and Moldovanu, 2001). Theorem I.1 provides an incentive-theoretical rationale for the utilitarian view in

[10] E.g., Jehiel, Moldovanu, and Stacchetti (1996, 1999), Jehiel and Moldovanu (2001), Goeree et al. (2005), Kucuksenel (2012), Lu (2012), and Tang and Sandholm (2012).

[11] Many of the studies on implementation in the presence of externalities are devoted to auction theory. Notice that the here derived propositions have only limited relevance for auctions, since I am concerned with budget balance while auction theory is typically concerned with revenue maximization. Moreover, I deal with continuous allocation rules whereas, in auctions, allocation rules are typically discrete.

[12] They refer to an allocation as 'efficient' if it maximizes the sum of payoffs *inclusive* of externalities.

[13] Bierbrauer et al. (2017) provide empirical evidence for the relevance of 'social-preference robust' implementation in the range of bilateral trade as well as income taxation. Bartling and Netzer (2016) follow a similar line for the design of auctions if bidders are privately informed about their spiteful preferences.

mechanism design theory, however complemented with the somewhat surprising qualification that, if externality assessments are private information, externalities must be *ignored* in order to achieve incentive compatibility *and* budget balance. Other foundations of utilitarianism have been provided on axiomatic grounds (e.g., Harsanyi, 1955, d'Aspremont and Gevers, 1977, and Maskin, 1978) and in the range of decision-making under ignorance (e.g., Maskin, 1979).

Theorem I.1 is bad news for the proponents of non-utilitarian measures of social welfare.[14] Examples for alternative concepts are the maximin-welfare measure of Rawls (1971), the CES-welfare measures proposed by Arrow (1973), and welfare measures that explicitly condition on indices of inequality (e.g., on the inequality index of Atkinson, 1970).[15] Theorem I.1 implies in particular that, in the presence of asymmetric information on externality assessments, incentive-compatible redistribution (beyond utilitarianism) comes at a price, embodied in the violation of budget balance.[16]

More generally, Theorems I.1 to I.2 suggest that theories of 'efficient' implementation depend critically on their underlying welfare judgments and might not pertain when introducing asymmetric information on agents' potentially irrational externality assessments. In this respect, the result also contributes to the growing field of behavioral mechanism design.[17] For instance, agents might not be able to fully process the information available (e.g., McFadden, 2009). Other agents might believe that there are externalities even though there are objectively none. Likewise, agents might be overly optimistic, or pessimistic, about how the well-being of others will affect themselves.[18] It seems plausible in all these cases that a social welfare measure should not condition on such 'behavioral' externality assessments, and that mechanisms designed to implement welfare maximizing allocations should be externality-robust.

[14]For critical reflections of utilitarianism see, e.g., Posner (1979) and Sen (1973, 1979).

[15]For a discussion of the CES-welfare measures see also Sen (1974).

[16]Saez and Stantcheva (2016), for instance, characterize optimal taxation under non-utilitarian social welfare measures—in absence of externalities though.

[17]E.g., Glazer and Rubinstein (1998), Cabrales and Serrano (2011), de Clippel (2014), Bierbrauer and Netzer (2016), and Bartling and Netzer (2016).

[18]In this respect, this study draws a mechanism design perspective on the 'tunnel effect' of Hirschman and Rothschild (1973).

## I.3. The Model

There is an interval $K = [k^{\min}, k^{\max}]$ of social alternatives, with $k^{\min} < k^{\max}$, and there are two agents, indexed by $i \in \{1,2\}$. The agent other than $i$ is denoted by $-i$. From alternative $k \in K$ and a monetary transfer $t_i \in \mathbb{R}$, agent $i$ gains a payoff

$$\pi_i(k, t_i \,|\, \theta_i) = \theta_i v_i(k) + h_i(k) + t_i, \tag{I.1}$$

where the functions $v_i : K \to [0, \infty)$ and $h_i : K \to \mathbb{R}$ are twice continuously differentiable and satisfy $\partial^2 \pi_i(k, t_i \,|\, \theta_i)/\partial k^2 < 0$ for all $i$, $k$, and $\theta_i > 0$; furthermore, either $dv_i/dk > 0$ for all $k$ and $i$, or $dv_i/dk < 0$ for all $k$ and $i$. The functions $v_i, h_i$ are common knowledge. Agent $i$'s *payoff type* $\theta_i$ is drawn from an interval $\Theta_i = (\theta_i^{\min}, \theta_i^{\max})$, with $0 \le \theta_i^{\min} < \theta_i^{\max}$. Payoff types are private information and are distributed according to a continuous density function $f_i > 0$. From the allocation of payoffs, agent $i$ gains utility

$$u_i(k, t_i, t_{-i}, \theta_{-i} \,|\, \theta_i, \delta_i) = \pi_i(k, t_i \,|\, \theta_i) + \delta_i \cdot \pi_{-i}(k, t_{-i} \,|\, \theta_{-i}), \tag{I.2}$$

where $i$'s *externality type* $\delta_i$ is drawn from an interval $\Delta_i = (\delta_i^{\min}, \delta_i^{\max}) \subset [-1, 1]$, with $\delta_i^{\min} < \delta_i^{\max}$. Externality types are private information and are distributed according to a continuous density function $g_i(\,\cdot\,|\,\theta_i) > 0$. That is, an agent's externality type may correlate with her payoff type, not perfectly though. Notice also that externality types take absolute values smaller than one, such that each agent's marginal utility from her own payoff exceeds her marginal (dis-)utility from her opponent's payoff.

Denote by $H_i$ the joint c.d.f. of agent $i$'s *type*, $(\theta_i, \delta_i)$. While types are private information, type distributions $H_i$ are common knowledge. Types are independent across agents; that is, $H_1$ and $H_2$ are stochastically independent.

Denote by $\Theta$ and $\Delta$, respectively, the Cartesian products $\Theta_1 \times \Theta_2$ and $\Delta_1 \times \Delta_2$, and let $\theta = (\theta_1, \theta_2)$ and $\delta = (\delta_1, \delta_2)$. For a random variable $X : \Theta \times \Delta \to \mathbb{R}$, denote by $\mathbb{E}_{\theta_i,\delta_i}\big[X(\theta,\delta)\big]$ the expected value of $X$ for given values of $\theta_{-i}$ and $\delta_{-i}$:[19]

$$\mathbb{E}_{\theta_i,\delta_i}\big[X(\theta,\delta)\big] = \int_{\theta_i^{\min}}^{\theta_i^{\max}} \left( \int_{\delta_i^{\min}}^{\delta_i^{\max}} X(\theta,\delta) g_i(\delta_i|\theta_i)\, d\delta_i \right) f_i(\theta_i)\, d\theta_i.$$

A *direct revelation mechanism* involves the agents in a strategic game. In this game, agents are asked to report their types truthfully.[20] Based on their reports, a social alternative will be implemented and transfers will be made. Specifically, the mechanism is defined by an *allocation rule* $k : \Theta \times \Delta \to K$ and a *transfer scheme* $T = (t_1, t_2) : \Theta \times \Delta \to \mathbb{R}^2$. In what follows, attention will be restricted to transfer schemes $T$ that are continuous on the externality-type space $\Delta$. An allocation rule $k$ is said to be *Bayesian implementable*, if there exists a transfer scheme $(t_1, t_2)$ such that both

$$\begin{aligned}
(\theta_1,\delta_1) &\in \arg\max_{\hat\theta_1,\hat\delta_1} \mathbb{E}_{\theta_2,\delta_2}\Big[u_1\big(k(\hat\theta_1,\hat\delta_1,\theta_2,\delta_2), t_1(\hat\theta_1,\hat\delta_1,\theta_2,\delta_2), t_2(\hat\theta_1,\hat\delta_1,\theta_2,\delta_2), \theta_2 \,\big|\, \theta_1,\delta_1\big)\Big],\\
(\theta_2,\delta_2) &\in \arg\max_{\hat\theta_2,\hat\delta_2} \mathbb{E}_{\theta_1,\delta_1}\Big[u_2\big(k(\theta_1,\delta_1,\hat\theta_2,\hat\delta_2), t_1(\theta_1,\delta_1,\hat\theta_2,\hat\delta_2), t_2(\theta_1,\delta_1,\hat\theta_2,\hat\delta_2), \theta_1 \,\big|\, \theta_2,\delta_2\big)\Big].
\end{aligned}$$

That is, truthful revelation maximizes each agent's interim expected utility provided the respective other agent reveals her type truthfully.

The mechanism is said to be *ex post budget-balanced* if the transfer scheme satisfies $t_1 + t_2 = 0$ for any realization of types, such that agents neither have to have access to an external source of money, nor that 'money is burned'.

The following two definitions restrict the domain of allocation rules to be considered in the next sections. For that purpose, define

$$\begin{aligned}
\pi_i(k\,|\,\theta_i) &= \theta_i v_i(k) + h_i(k), \quad \text{and}\\
u_i(k,\theta_{-i}\,|\,\theta_i,\delta_i) &= \pi_i(k\,|\,\theta_i) + \delta_i \pi_{-i}(k\,|\,\theta_{-i}),
\end{aligned}$$

[19] Likewise, denote by $\mathbb{E}_{\theta_i}\big[Y(\theta)\big]$ the expected value of $Y : \Theta \to \mathbb{R}$ for a given value of $\theta_{-i}$.

[20] By the *revelation principle*, which applies to the present setup (Myerson, 1979), there is no loss of generality in identifying message sets, from which agents draw their reports, with agents' type sets.

and denote by $\mathrm{sgn}:\mathbb{R}\to\{-1,0,1\}$ the *sign function*.[21]

**Definition I.1 (Sensitivity)**

*Let $W:\mathbb{R}^4\to\mathbb{R}$ be twice partially continuously differentiable, and let $V:K\to\mathbb{R}$, $V(k)=W\big(\pi_1(k\,|\,\theta_1),\,\delta_1\pi_2(k\,|\,\theta_2),\,\pi_2(k\,|\,\theta_2),\,\delta_2\pi_1(k\,|\,\theta_1)\big)$.[22] $W$ is said to be a sensitive social welfare measure if it has the following properties.*

*(i)* $\partial W(\pi_1,\delta_1\pi_2,\pi_2,\delta_2\pi_1)/\partial\pi_i>0$ *for each* $i\in\{1,2\}$.

*(ii) Pareto property: If there exist $k_1,k_2\in K$ and $i\in\{1,2\}$ such that $u_i(k_1,\theta_{-i}\,|\,\theta_i,\delta_i)>u_i(k_2,\theta_{-i}\,|\,\theta_i,\delta_i)$ and $u_{-i}(k_1,\theta_i\,|\,\theta_{-i},\delta_{-i})\geq u_{-i}(k_2,\theta_i\,|\,\theta_{-i},\delta_{-i})$, then $V(k_1)>V(k_2)$.*

*(iii) There exists a unique partially continuously differentiable allocation rule $k^*:\Theta\times\Delta\to K$ such that $k^*(\theta,\delta)=\arg\max_{k\in K}V(k)$,*

$$1 = \mathrm{sgn}\left(\frac{\partial v_1(k^*)}{\partial\theta_2}\right)\cdot\mathrm{sgn}\left(\frac{\partial v_2(k^*)}{\partial\theta_1}\right),\quad \textit{and} \tag{I.3}$$

$$0 = \mathrm{sgn}\left(\frac{\partial\pi_1(k^*\,|\,\theta_1)}{\partial\delta_2}\right)-\mathrm{sgn}\left(\frac{\partial\pi_2(k^*\,|\,\theta_2)}{\partial\delta_1}\right). \tag{I.4}$$

*The allocation rule $k^*$ is said to be sensitive.*

Whether a function qualifies as a sensitive social welfare measure is context-dependent, since the above conditions involve the functions $v_i$ and $h_i$. A sensitive social welfare measure $V(k)$ accounts separately for private payoffs, $\pi_i(k\,|\,\theta_i)$, and externalities, $\delta_i\pi_{-i}(k\,|\,\theta_{-i})$. This serves the purpose of clearly isolating the extent to which 'efficient' allocation rules may condition on externality assessments if they are to be Bayesian implemented in a budget-balanced way.

By condition (i), a marginal increase in one agent's private payoff contributes to social welfare. Conditions (ii) and (iii), jointly, ensure that the allocation rule *unambiguously* specifies *some* allocation on the ex post Pareto frontier. According to equations (I.1) and (I.2), full ex post Pareto efficiency is realized if, in addition, transfers are budget-balanced.

[21]For $x\in\mathbb{R}$, the sign of $x$ is defined as $\mathrm{sgn}(x)=1$ for $x>0$, $\mathrm{sgn}(x)=-1$ for $x<0$, and $\mathrm{sgn}(0)=0$.

[22]This specification of a welfare measure with regard to the choice of $k$ is without loss of generality as it allows for taking private payoffs, $\pi_i$, and externality types, $\delta_i$, as independent variables. For instance, $V(k)=(1+\delta_1^2)\pi_2+(1+\delta_2^2)\pi_1$ can be written as $V(k)=(\pi_1)+(\pi_2)+\frac{(\delta_2\pi_1)}{(\pi_1)}(\delta_2\pi_1)+\frac{(\delta_1\pi_2)}{(\pi_2)}(\delta_1\pi_2)$.

Identities (I.3) and (I.4) are symmetry assumptions. Identity (I.3) requires that the effect of an increase in agent $i$'s payoff type on agent $-i$'s payoff, exclusive of $h_{-i}(k^*)$, is similar for all agents. As the functions $v_i$ are assumed to be either strictly increasing or strictly decreasing, and since $\partial v_i(k^*)/\partial\theta_{-i} = (dv_i(k^*)/dk)(\partial k^*/\partial\theta_{-i})$, equation (I.3) requires in particular that a sensitive allocation rule is either strictly increasing or strictly decreasing in each agent's payoff type. In this respect, it responds *sensitively* to changes in agents' payoff characteristics.[23] Finally, identity (I.4) requires that the effect of an increase in one agent's externality type on the other agent's payoff is similar for all agents.

Several "classic" social welfare measures qualify as sensitive.[24]

**Proposition I.1** *With notation as in Definition I.1, each of the following social welfare measures* $W : \mathbb{R}^4 \to \mathbb{R}$ *is sensitive if the economic environment is such that* $W$ *induces a unique partially continuously differentiable function* $k^*(\theta,\delta) = \arg\max_{k\in K} V(k)$ *satisfying* $\partial k^*/\partial\theta_i \neq 0$ *for all* $(\theta,\delta)\in\Theta\times\Delta$ *and all* $i\in\{1,2\}$.

*(i) Externality-ignoring utilitarianism:* $W = \pi_1(k\,|\,\theta_1) + \pi_2(k\,|\,\theta_2)$.

*(ii) Externality-sensitive utilitarianism:* $W = u_1(k,\theta_2\,|\,\theta_1,\delta_1) + u_2(k,\theta_1\,|\,\theta_2,\delta_2)$.

*If the economic environment is restricted to* $h_i \equiv 0$ *and* $\Delta_i \subset [0,1)$ *for all* $i\in\{1,2\}$, *then the following social welfare measures are sensitive.*

*(iii) "Social utility weights", inclusive of externalities:*

$$W = \alpha_1 u_1(k,\theta_2\,|\,\theta_1,\delta_1) + \alpha_2 u_2(k,\theta_1\,|\,\theta_2,\delta_2), \text{ with } \alpha_1,\alpha_2 > 0.$$

*(iv) The Nash product, inclusive of externalities:*

$$W = u_1(k,\theta_2\,|\,\theta_1,\delta_1)\cdot u_2(k,\theta_1\,|\,\theta_2,\delta_2).$$

---

[23] Notice also that condition (iii) of Definition I.1 requires the economic environment as well as a sensitive social welfare measure to allow for *interior* solutions to $\max_{k\in K} V(k)$. Hence, $k^*$ must satisfy the first-order condition $dV(k^*(\theta,\delta))/dk = 0$ and the second-order condition $d^2V(k^*(\theta,\delta))/dk^2 < 0$ for each $(\theta,\delta)\in\Theta\times\Delta$.

[24] Notice that condition (I.4) of Definition I.1 precludes the dictatorial social welfare measure $V(k) = u_i(k,\theta_{-i}\,|\,\theta_i,\delta_i)$ from being sensitive, since then $\partial\pi_i(k^*\,|\,\theta_i)/\partial\delta_{-i} = 0$, whereas $\partial\pi_{-i}(k^*\,|\,\theta_{-i})/\partial\delta_i \neq 0$. Notice further that externality-ignoring discriminatory utilitarianism of the form $W = \alpha_1\pi_1(k\,|\,\theta_1) + \alpha_2\pi_2(k\,|\,\theta_2)$, with $\alpha_1,\alpha_2 > 0$ and $\alpha_1 \neq \alpha_2$, satisfies all the conditions of Definition I.1 but might not have the Pareto property.

*(v) CES-welfare, inclusive of externalities:*

$$W = \left[(u_1(k,\theta_2\,|\,\theta_1,\delta_1))^{-\rho} + (u_2(k,\theta_1\,|\,\theta_2,\delta_2))^{-\rho}\right]^{-\frac{1}{\rho}},\ \textit{with}\ \rho \in (-1,\infty)\setminus\{0\}.$$

**Proof.** Externality-ignoring utilitarianism will be addressed separately in Proposition I.2. Proofs are straightforward for (ii) and (iii) and are thus omitted. See the Appendix for (iv) and (v). ∎

By means of the next definition, attention will be further restricted to those Bayesian implementable allocation rules that do not condition on (moments of) type distributions.

**Definition I.2 (Strong Bayesian implementability)**
*An allocation rule $k^* : \Theta \times \Delta \to K$ is said to be strongly Bayesian implementable if it is Bayesian implementable for any set of (non-degenerate) type distributions, $\{F_1, G_1, F_2, G_2\}$.*

Strong Bayesian implementability is critical to the results obtained below.[25] It does not require the mechanism as a whole to be independent from type distributions. It rather makes a qualitative distinction between 'means' (the transfer scheme) and 'ends' (the allocation rule). The social welfare judgment inherent to this concept is that ex post allocations ought not depend on what agents' types *could have been* but only on what agents' types *are ex post*.[26]

### I.3.1. Altruism, Spite, and Status Considerations

Evidently, the model captures the linear conceptions of altruism and spite when interpreting externality types as the intensity of altruism or spite. It also captures linear conceptions of preferences for status:[27] Suppose the allocation of payoffs, $\pi_1$ and $\pi_2$, yields agent $i$ a utility level of $u_i = \pi_i + \sigma_i(\pi_i - \pi_{-i})$, with $\sigma_i > 0$ determining $i$'s preference for status. Maximizing $u_i$ is then equivalent to maximizing $\hat{u}_i = u_i/(1+\sigma_i) = \pi_i + \delta_i \pi_{-i}$, with externality type $\delta_i = -\sigma_i/(1+\sigma_i) \in (-1, 0)$.

[25] Strong Bayesian implementability should not be confused with notions of 'robust' implementation in the manner of Bergemann and Morris (2009, 2013).
[26] An example of a social welfare measure that *does* condition on type distributions is the *generalized Nash product* of Harsanyi and Selten (1972).
[27] See, e.g., Boskin and Sheshinski (1978) and Bisin and Verdier (1998). By the same token, the model captures linear versions of interdependent utilities in the manner of Hirschman and Rothschild (1973).

## I.4. The Incentive Costs of Welfare Judgments

This section proves the following theorem (employing Propositions I.2 to I.4) and discusses it from various angles (through Theorems I.2 and I.3).

**Theorem I.1** *A sensitive allocation rule $k^* : \Theta \times \Delta \to K$ can be strongly Bayesian implemented with an ex post budget-balanced mechanism if and only if it maximizes the sum of private payoffs exclusive of externalities: $k^*(\theta, \delta) = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$ for all $(\theta, \delta)$; in particular, $k^*$ is independent from externality types: $k^* = k^*|_\Theta$.*

*Any mechanism that (ordinarily) Bayesian implements $k^*(\theta) = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$ is of AGV-type: For reported types $(\hat{\theta}, \hat{\delta}) \in \Theta \times \Delta$, transfers are given by*

$$t_1(\hat{\theta}, \hat{\delta}) = \mathbb{E}_{\theta_2}\big[\pi_2(k^*(\hat{\theta}_1, \theta_2) \,|\, \theta_2)\big] - \mathbb{E}_{\theta_1}\big[\pi_1(k^*(\theta_1, \hat{\theta}_2) \,|\, \theta_1)\big] + s(\hat{\theta}, \hat{\delta}), \tag{I.5}$$

$$t_2(\hat{\theta}, \hat{\delta}) = \mathbb{E}_{\theta_1}\big[\pi_1(k^*(\theta_1, \hat{\theta}_2) \,|\, \theta_1)\big] - \mathbb{E}_{\theta_2}\big[\pi_2(k^*(\hat{\theta}_1, \theta_2) \,|\, \theta_2)\big] - s(\hat{\theta}, \hat{\delta}), \tag{I.6}$$

*where $s : \Theta \times \Delta \to \mathbb{R}$ must be chosen such that $\mathbb{E}_{\theta_{-i}, \delta_{-i}}[s(\theta, \delta)]$ is constant on $\Theta_i \times \Delta_i$ for each $i \in \{1, 2\}$.*[28]

By Theorem I.1, Bayesian implementation of a social welfare judgment inconsistent with externality-ignoring utilitarianism violates budget balance and thus entails incentive costs.

In the following, I refer to the mechanisms specified by Theorem I.1 as *AGV-type mechanisms* (after Arrow, 1979, and d'Aspremont and Gérard-Varet, 1979). Notice that, *ex interim*, AGV-type mechanisms leave externality assessments strategically inoperative. If the distribution of externality types is not common knowledge, one can let $s = 0$.

The sufficient conditions of Theorem I.1 as well as the sensitivity of externality-ignoring utilitarianism are to be addressed first.

**Proposition I.2** *Suppose the allocation rule $k^* : \Theta \to K$ is partially continuously differentiable and satisfies $k^*(\theta) = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$ and $\partial k^* / \partial \theta_i \neq 0$ for all*

[28]Such functions $s$ can be smooth and non-constant; for instance, $s(\theta, \delta) = (\theta_1 - \mathbb{E}_{\theta_1}[\theta_1])(\theta_2 - \mathbb{E}_{\theta_2}[\theta_2]) + (\delta_1 - \mathbb{E}_{\delta_1}[\delta_1])(\delta_2 - \mathbb{E}_{\delta_2}[\delta_2])$.

$\theta \in \Theta$. *Then* $k^*$ *is sensitive and can be strongly Bayesian implemented with the ex post budget-balanced AGV-type mechanisms.*[29]

**Proof.** In order to prove the sensitivity of $k^*$, it suffices to show that $k^*$ does specify Pareto-efficient allocations. (Verification of the remaining properties of a sensitive allocation rule follows the lines of the proof of Proposition I.1(ii).)

Suppose there exists an allocation $k'(\theta, \delta)$ that, for some types $(\theta, \delta)$, Pareto-improves upon $k^*(\theta)$. Since $\pi_i(k \,|\, \theta_i)$ is concave, $\pi_1(k'(\theta,\delta) \,|\, \theta_1) + \pi_2(k'(\theta,\delta) \,|\, \theta_2) < \pi_1(k^*(\theta) \,|\, \theta_1) + \pi_2(k^*(\theta) \,|\, \theta_2)$. Suppose agent 1 suffers the (weakly) greater loss in private payoffs. Then the differences $d_i = \pi_i(k^*(\theta) \,|\, \theta_i) - \pi_i(k'(\theta,\delta) \,|\, \theta_i)$ satisfy $d_1 > 0$ and $d_1 \geq d_2 > -d_1$. Since $\delta_1 \in \Delta_1 \subset (-1, 1)$,

$$u_1(k'(\theta,\delta), \theta_2 \,|\, \theta_1, \delta_1) - u_1(k^*(\theta), \theta_2 \,|\, \theta_1, \delta_1) = -(d_1 + \delta_1 d_2) < 0.$$

Hence, agent 1 is worse of under $k'(\theta, \delta)$ than under $k^*(\theta)$; a contradiction.

Under AGV-type mechanisms in the manner of Theorem I.1 (which, evidently, are ex post budget-balanced), and under the assumption that agent 2 reveals her type $(\theta_2, \delta_2)$ truthfully, agent 1 chooses $(\hat{\theta}_1, \hat{\delta}_1)$ so as to maximize her interim expected utility. Without loss of generality, normalize $s(\hat{\theta}, \hat{\delta}) = 0$. By equations (I.5) and (I.6),

$$\begin{aligned}
&\mathbb{E}_{\theta_2,\delta_2}\Big[u_1\big(k^*(\hat{\theta}_1,\theta_2), t_1(\hat{\theta}_1,\theta_2,\hat{\delta}), t_2(\hat{\theta}_1,\theta_2,\hat{\delta}), \theta_2 \,\big|\, \theta_1, \delta_1\big)\Big] \\
= \;& \mathbb{E}_{\theta_2}\Big[\big[\pi_1(k^*(\hat{\theta}_1,\theta_2) \,|\, \theta_1) + t_1(\hat{\theta}_1,\theta_2,\hat{\delta})\big] + \delta_1 \cdot \big[\pi_2(k^*(\hat{\theta}_1,\theta_2) \,|\, \theta_2) + t_2(\hat{\theta}_1,\theta_2,\hat{\delta})\big]\Big] \\
= \;& \mathbb{E}_{\theta_2}\big[\pi_1(k^*(\hat{\theta}_1,\theta_2) \,|\, \theta_1) + \pi_2(k^*(\hat{\theta}_1,\theta_2) \,|\, \theta_2)\big] - (1-\delta_1)\mathbb{E}_{\theta_1,\theta_2}\big[\pi_1(k^*(\theta_1,\theta_2) \,|\, \theta_1)\big],
\end{aligned}$$

where the second term in the last line is independent from $\hat{\theta}_1$. Suppose truthfully reporting $\theta_1$ is strictly inferior to some report $\hat{\theta}_1 \neq \theta_1$. Then there must exist some $\theta_2$ such that

$$\pi_1(k^*(\hat{\theta}_1,\theta_2) \,|\, \theta_1) + \pi_2(k^*(\hat{\theta}_1,\theta_2) \,|\, \theta_2) > \pi_1(k^*(\theta_1,\theta_2) \,|\, \theta_1) + \pi_2(k^*(\theta_1,\theta_2) \,|\, \theta_2),$$

[29] That AGV-type mechanisms are Bayesian incentive-compatible for other-regarding, spiteful agents has been shown earlier by Bartling and Netzer (2016).

which contradicts the definition of $k^*$. Hence, agent 1 has no incentive to misreport her payoff type. Obviously, she has no incentive to misreport her externality type. By symmetry, agent 2 cannot do better than reporting $(\theta_2, \delta_2)$. As the argument holds for any set of type distributions, AGV-type mechanisms *strongly* Bayesian implement $k^*$. ∎

As becomes clear from the proof of Proposition I.2, the model assumption that each agent's marginal utility from her own payoff exceeds her marginal (dis-)utility from her opponent's payoff is indeed critical. For 'excessive' externalities, $|\delta_i| > 1$, externality-ignoring utilitarianism in the manner of Theorem I.1 will not generally lead to ex post Pareto-efficient allocations. However, I have presented several examples of economic environments for which the assumption of 'moderate' externalities, $|\delta_i| < 1$, is reasonable.

Evidently, Proposition I.2 holds for more general (e.g., multi-dimensional) sets of *payoff types* - a property of the AGV-mechanism which is well-known for environments without externality assessments (see, e.g., Mas-Colell, Whinston, and Green, 1995, ch.23).

The following two propositions give proof of the necessary conditions of Theorem I.1. These propositions successively constrain the domain of sensitive allocation rules and budget-balanced transfer schemes that can be strongly Bayesian implemented. They stipulate externality robustness in the sense that externality assessments are left inoperative from a strategic point of view. The following Lemma eases the exposition.

**Lemma I.1** *Suppose the partially differentiable allocation rule $k^* : \Theta \times \Delta \to K$ is strongly Bayesian implementable with an ex post budget-balanced mechanism. Then $k^*$ satisfies*

$$(1-\delta_i)\frac{\partial v_i(k^*(\theta,\delta))}{\partial \delta_i} = \left[\frac{d\pi_i(k^*(\theta,\delta)\,|\,\theta_i)}{dk} + \frac{d\pi_{-i}(k^*(\theta,\delta)\,|\,\theta_{-i})}{dk}\right]\frac{\partial k^*(\theta,\delta)}{\partial \theta_i} \tag{I.7}$$

*for all $(\theta,\delta) \in \Theta \times \Delta$ and all $i \in \{1,2\}$. If $k^*$ is independent from externality types, $k^* = k^*|_\Theta$, then $k^*$ is (ordinarily) Bayesian implementable in a budget-balanced way only if the transfer to each agent $i$ satisfies*

$$\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_i(\theta,\delta)\big] = \alpha_i + \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_{-i}(k^*(\theta,\delta)\,|\,\theta_{-i})\big] \tag{I.8}$$

*for all* $(\theta_i, \delta_i) \in \Theta_i \times \Delta_i$ *and some constant* $\alpha_i \in \mathbb{R}$.

**Proof.** See the Appendix. ■

In light of the second part of Lemma I.1, the following proposition implies that the desired mechanism may not condition on externality types, such that externality assessments are not directly strategically operative.

**Proposition I.3** *A sensitive allocation rule* $k^* : \Theta \times \Delta \to K$ *is strongly Bayesian implementable with an ex post budget-balanced mechanism only if it is independent from externality types:* $k^* = k^*|_\Theta$.

**Proof.** Let $k^* : \Theta \times \Delta \to K$ be the sensitive allocation rule that corresponds to a sensitive social welfare measure $W : \mathbb{R}^4 \to \mathbb{R}$. Ease notation by writing $k^* = k^*(\theta, \delta)$. It has to be shown that $\partial k^*/\partial \delta_i = 0$ for all $(\theta, \delta) \in \Theta \times \Delta$ and all $i \in \{1, 2\}$.

For $x \in \mathbb{R}^4$ and $j = \{1, ..., 4\}$, write $W_j(x) = \partial W(x)/\partial x_j$, and define

$$W_j = W_j\big(\pi_1(k^* \,|\, \theta_1), \delta_1 \pi_2(k^* \,|\, \theta_2), \pi_2(k^* \,|\, \theta_2), \delta_2 \pi_1(k^* \,|\, \theta_1)\big). \tag{I.9}$$

Then the conditions of Definition I.1 imply that $k^*$ satisfies the FOC

$$0 = \frac{dV(k^*)}{dk} = [W_1 + \delta_2 W_4] \frac{d\pi_1(k^* \,|\, \theta_1)}{dk} + [W_3 + \delta_1 W_2] \frac{d\pi_2(k^* \,|\, \theta_2)}{dk}, \tag{I.10}$$

where $W_1 + \delta_2 W_4 = \partial W/\partial \pi_1 > 0$ and $W_3 + \delta_1 W_2 = \partial W/\partial \pi_2 > 0$ by Definition I.1(i).

By Lemma I.1, $k^*$ satisfies also

$$(1 - \delta_1) \frac{\partial v_1(k^*)}{\partial \delta_1} = \left[ \frac{d\pi_1(k^* \,|\, \theta_1)}{dk} + \frac{d\pi_2(k^* \,|\, \theta_2)}{dk} \right] \frac{\partial k^*}{\partial \theta_1}, \tag{I.11}$$

$$(1 - \delta_2) \frac{\partial v_2(k^*)}{\partial \delta_2} = \left[ \frac{d\pi_1(k^* \,|\, \theta_1)}{dk} + \frac{d\pi_2(k^* \,|\, \theta_2)}{dk} \right] \frac{\partial k^*}{\partial \theta_2}. \tag{I.12}$$

Substituting (I.10) into (I.11) and (I.12) yields

$$(1-\delta_1)\frac{\partial v_1(k^*)}{\partial \delta_1} = \left[1 - \frac{W_1+\delta_2 W_4}{W_3+\delta_1 W_2}\right]\frac{d\pi_1(k^*\,|\,\theta_1)}{dk}\frac{\partial k^*}{\partial \theta_1}, \tag{I.13}$$

$$(1-\delta_2)\frac{\partial v_2(k^*)}{\partial \delta_2} = \left[1 - \frac{W_3+\delta_1 W_2}{W_1+\delta_2 W_4}\right]\frac{d\pi_2(k^*\,|\,\theta_2)}{dk}\frac{\partial k^*}{\partial \theta_2}. \tag{I.14}$$

On the other hand, as $\partial k^*/\partial\theta_i \neq 0$ by Definition I.1(iii), identities (I.11) and (I.12) jointly imply that

$$(1-\delta_1)\frac{\partial v_1(k^*)}{\partial \delta_1}\frac{\partial k^*}{\partial \theta_2} = (1-\delta_2)\frac{\partial v_2(k^*)}{\partial \delta_2}\frac{\partial k^*}{\partial \theta_1}. \tag{I.15}$$

As $\delta_i < 1$ and $dv_i/dk \neq 0$ by assumption, identity (I.15) implies that $\partial k^*/\partial\delta_1 = 0$ if and only if $\partial k^*/\partial\delta_2 = 0$.

Suppose $\partial k^*(\theta,\delta)/\partial\delta_i \neq 0$ for some $(\theta,\delta)$ and all $i$. Then each of the factors on the right-hand sides of (I.13) and (I.14) is non-zero. In this case, (I.13) and (I.14) imply that

$$\begin{aligned}\frac{(W_3+\delta_1 W_2)(1-\delta_1)\frac{\partial v_1(k^*)}{\partial\delta_1}}{\frac{d\pi_1(k^*\,|\,\theta_1)}{dk}\frac{\partial k^*}{\partial\theta_1}} &= [(W_3+\delta_1 W_2)-(W_1+\delta_2 W_4)] \\ &= -[(W_1+\delta_2 W_4)-(W_3+\delta_1 W_2)] \\ &= -\frac{(W_1+\delta_2 W_4)(1-\delta_2)\frac{\partial v_2(k^*)}{\partial\delta_2}}{\frac{d\pi_2(k^*\,|\,\theta_2)}{dk}\frac{\partial k^*}{\partial\theta_2}}.\end{aligned} \tag{I.16}$$

Rearranging (I.16), while writing $\frac{\partial v_i(k^*)}{\partial\delta_i} = \frac{dv_i(k^*)}{dk}\frac{\partial k^*}{\partial\delta_i}$, yields the identity

$$\begin{aligned}&(W_3+\delta_1 W_2)(1-\delta_1)\frac{dv_1(k^*)}{dk}\frac{\partial k^*}{\partial\delta_1}\frac{d\pi_2(k^*\,|\,\theta_2)}{dk}\frac{\partial k^*}{\partial\theta_2} \\ = \; &-(W_1+\delta_2 W_4)(1-\delta_2)\frac{dv_2(k^*)}{dk}\frac{\partial k^*}{\partial\delta_2}\frac{d\pi_1(k^*\,|\,\theta_1)}{dk}\frac{\partial k^*}{\partial\theta_1}.\end{aligned} \tag{I.17}$$

Since the terms $(W_1+\delta_2 W_4)$, $(W_3+\delta_1 W_2)$, and $(1-\delta_i)$ are positive, application of the *sign function* to each side of identity (I.17) yields

$$\operatorname{sgn}\left(\frac{d\pi_2(k^*\,|\,\theta_2)}{dk}\frac{\partial k^*}{\partial\delta_1}\frac{dv_1(k^*)}{dk}\frac{\partial k^*}{\partial\theta_2}\right) = -\operatorname{sgn}\left(\frac{d\pi_1(k^*\,|\,\theta_1)}{dk}\frac{\partial k^*}{\partial\delta_2}\frac{dv_2(k^*)}{dk}\frac{\partial k^*}{\partial\theta_1}\right). \tag{I.18}$$

By Definition I.1(iii), $\operatorname{sgn}(\partial v_1(k^*)/\theta_2) \cdot \operatorname{sgn}(\partial v_2(k^*)/\theta_1) = 1$, such that (I.18) can only hold if

$$\operatorname{sgn}\left(\frac{\partial \pi_2(k^* \mid \theta_2)}{\partial \delta_1}\right) = -\operatorname{sgn}\left(\frac{\partial \pi_1(k^* \mid \theta_1)}{\partial \delta_2}\right). \tag{I.19}$$

Equation (I.19) contradicts identity (I.4) of Definition I.1, unless $\partial \pi_i(k^* \mid \theta_i)/\partial \delta_{-i} = 0$ for all $i$. Suppose $\partial \pi_1(k^* \mid \theta_1)/\partial \delta_2 = 0$; then multiplying (I.13) with $\partial k^*(\theta, \delta)/\partial \delta_2$ implies that

$$(1 - \delta_1)\frac{\partial v_1(k^*)}{\partial \delta_1}\frac{\partial k^*(\theta, \delta)}{\partial \delta_2} = 0. \tag{I.20}$$

As $\delta_i < 1$ and $dv_i/dk \neq 0$ by assumption, (I.20) yields $\frac{\partial k^*(\theta,\delta)}{\partial \delta_1}\frac{\partial k^*(\theta,\delta)}{\partial \delta_2} = 0$, such that $\frac{\partial k^*(\theta,\delta)}{\partial \delta_1} = 0 = \frac{\partial k^*(\theta,\delta)}{\partial \delta_2}$ due to (I.15) and the reasoning thereafter. Hence, $k^* = k^*|_\Theta$. ■

Externality assessments might indirectly become strategically operative if the allocation rule, even if independent from externality types, unfolds redistributive effects (beyond Benthamite utilitarianism). The next proposition states that strong, budget-balanced Bayesian implementation of a sensitive allocation rule is only feasible if the underlying social welfare measure treats agents' private payoffs as perfect substitutes.

**Proposition I.4** *A sensitive allocation rule $k^* : \Theta \to K$, which is independent from externality types, is strongly Bayesian implementable with an ex post budget-balanced mechanism only if $k^*(\theta) = \arg\max_{k \in K} \pi_1(k \mid \theta_1) + \pi_2(k \mid \theta_2)$ for all $\theta \in \Theta$; any mechanism that (ordinarily) Bayesian implements this allocation rule is necessarily of AGV-type.*

**Proof.** If $k^*$ is independent from externality types, identity (I.7) of Lemma I.1 becomes

$$0 = \left[\frac{d\pi_i(k^*(\theta) \mid \theta_i)}{dk} + \frac{d\pi_{-i}(k^*(\theta) \mid \theta_{-i})}{dk}\right]\frac{\partial k^*(\theta)}{\partial \theta_i}.$$

By Definition I.1(iii), either $\partial k^*/\partial \theta_i > 0$ for all $\theta_i$, or $\partial k^*/\partial \theta_i < 0$ for all $\theta_i$. Hence, $k^*(\theta) = \arg\max_{k \in K} \pi_1(k \mid \theta_1) + \pi_2(k \mid \theta_2)$ for all $\theta \in \Theta$.

Suppose there exists a budget-balanced transfer scheme $T^* = (t_1^*, t_2^*) : \Theta \times \Delta \to \mathbb{R}^2$ that Bayesian implements $k^*(\theta) = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$. Notice that one can always write

$$t_1^*(\hat{\theta}, \hat{\delta}) = \mathbb{E}_{\theta_2}\big[\pi_2(k^*(\hat{\theta}_1, \theta_2) \,|\, \theta_2)\big] - \mathbb{E}_{\theta_1}\big[\pi_1(k^*(\theta_1, \hat{\theta}_2) \,|\, \theta_1)\big] + s_1(\hat{\theta}, \hat{\delta}), \quad \text{(I.21)}$$

$$t_2^*(\hat{\theta}, \hat{\delta}) = \mathbb{E}_{\theta_1}\big[\pi_1(k^*(\theta_1, \hat{\theta}_2) \,|\, \theta_1)\big] - \mathbb{E}_{\theta_2}\big[\pi_2(k^*(\hat{\theta}_1, \theta_2) \,|\, \theta_2)\big] + s_2(\hat{\theta}, \hat{\delta}), \quad \text{(I.22)}$$

for appropriate functions $s_1, s_2 : \Theta \times \Delta \to \mathbb{R}$ that satisfy $s_1 + s_2 = 0$ on $\Theta \times \Delta$. But then, for each $i \in \{1, 2\}$ and all $(\theta_i, \delta_i) \in \Theta_i \times \Delta_i$,

$$\mathbb{E}_{\theta_{-i}, \delta_{-i}}\big[t_i^*(\theta, \delta)\big] = \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^*(\theta) \,|\, \theta_{-i})\big] - \mathbb{E}_{\theta_i, \theta_{-i}}\big[\pi_i(k^*(\theta) \,|\, \theta_i)\big] + \mathbb{E}_{\theta_{-i}, \delta_{-i}}\big[s_i(\theta, \delta)\big]. \quad \text{(I.23)}$$

On the other hand, for $k^* : \Theta \to K$, Lemma I.1 states that

$$\mathbb{E}_{\theta_{-i}, \delta_{-i}}\big[t_i^*(\theta, \delta)\big] = \alpha_i + \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^*(\theta) \,|\, \theta_{-i})\big] \quad \text{(I.24)}$$

for all $(\theta_i, \delta_i) \in \Theta_i \times \Delta_i$ and some constant $\alpha_i \in \mathbb{R}$. Jointly, identities (I.23) and (I.24) imply that $\mathbb{E}_{\theta_{-i}, \delta_{-i}}\big[s_i(\theta, \delta)\big] = \alpha_i + \mathbb{E}_{\theta_i, \theta_{-i}}\big[\pi_i(k^*(\theta) \,|\, \theta_i)\big]$ for all $(\theta_i, \delta_i)$, so that $\mathbb{E}_{\theta_{-i}, \delta_{-i}}\big[s_i(\theta, \delta)\big]$ must be constant on $\Theta_i \times \Delta_i$. Hence, the mechanism $(k^*, T^*)$ is of *AGV-type*. ■

Propositions I.2 to I.4 give proof of Theorem I.1. The next result emphasizes the critical role of budget balance when it comes to Bayesian implementation of social welfare judgments in the presence of asymmetric information about agents' externality assessments.

**Theorem I.2** *If one waives budget balance, any twice continuously differentiable allocation rule $k^* : \Theta \times \Delta \to K$ satisfying $\min_{\theta_i, \delta_i} \frac{\partial}{\partial \theta_i} \mathbb{E}_{\theta_{-i}, \delta_{-i}}\big[v_i(k^*(\theta, \delta))\big] > 0$ for all $i$ can be strongly Bayesian implemented. If $k^* = k^*|_\Theta$, then $\frac{\partial}{\partial \theta_i} \mathbb{E}_{\theta_{-i}}\big[v_i(k^*(\theta, \delta))\big] \geq 0$ is sufficient.*

**Proof.** The straight forward proof for allocation rules satisfying $k^* = k^*|_\Theta$ is put to the Appendix. Let $k^* : \Theta \times \Delta \to K$, with $k^* \neq k^*|_\Theta$, be a twice continuously differentiable allocation rule satisfying $\beta_i > 0$ for $\beta_i = \min_{\theta_i,\delta_i} \frac{\partial}{\partial\theta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\theta,\delta))\big]$. For functions $p_i : \Delta_i \to \mathbb{R}$ define the transfer scheme $T^* = (t_1^*, t_2^*) : \Theta \times \Delta \to \mathbb{R}^2$ by

$$\begin{aligned} t_i^*(\hat{\theta},\hat{\delta}) \quad = \quad & p_i(\hat{\delta}_i) - \hat{\delta}_i \frac{\partial p_i(\hat{\delta}_i)}{\partial \hat{\delta}_i} + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(s,\hat{\theta}_{-i},\hat{\delta}_i,\hat{\delta}_{-i}))\big]ds \\ & + \frac{\partial p_{-i}(\hat{\delta}_{-i})}{\partial \hat{\delta}_{-i}} + \frac{\partial}{\partial \hat{\delta}_{-i}} \int_{\theta_{-i}^{\min}}^{\hat{\theta}_{-i}} \mathbb{E}_{\theta_i,\delta_i}\big[v_{-i}(k^*(\hat{\theta}_i,s,\hat{\delta}_i,\hat{\delta}_{-i}))\big]ds \\ & - \hat{\delta}_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(s,\hat{\theta}_{-i},\hat{\delta}_i,\hat{\delta}_{-i}))\big]ds \\ & - \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_i(k^*(\hat{\theta},\hat{\delta})\,|\,\hat{\theta}_i)\big] - \mathbb{E}_{\theta_i,\delta_i}\big[\pi_i(k^*(\hat{\theta},\hat{\delta})\,|\,\hat{\theta}_i)\big]. \end{aligned} \tag{I.25}$$

Then $T^*$ strongly Bayesian implements $k^*$ if the functions $p_i$ are chosen such that the following condition holds for all $(\theta_i,\delta_i)$ and all $i$:

$$\frac{\Big[\frac{\partial}{\partial\delta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\theta,\delta))\big]\Big]^2}{\frac{\partial}{\partial\theta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\theta,\delta))\big]} < \frac{\partial^2}{\partial\delta_i^2}\left[p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(s,\theta_{-i},\delta_i,\delta_{-i}))\big]ds\right]. \tag{I.26}$$

For example, one can choose $p_i(\delta_i) = \frac{1}{2}c_i\delta_i^2$, with

$$c_i = \gamma_i - \min_{\theta_i,\delta_i} \frac{\partial^2}{\partial\delta_i^2} \int_{\theta_i^{\min}}^{\theta_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(s,\theta_{-i},\delta_i,\delta_{-i}))\big]ds \tag{I.27}$$

for some constant $\gamma_i$ satisfying $\beta_i \cdot \gamma_i > \max_{\theta_i,\delta_i}\Big[\frac{\partial}{\partial\delta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\theta,\delta))\big]\Big]^2$.[30] For an extensive proof of this claim as well as a derivation of $T^*$, see the Appendix. ■

Notice that the assumption of $\min_{\theta_i,\delta_i} \frac{\partial}{\partial\theta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\theta,\delta))\big] > 0$ in Theorem I.2 is fairly weak; as implied by condition (A.16) in the proof of Lemma I.1, any Bayesian implementable allocation rule $k^*$ necessarily satisfies $\frac{\partial}{\partial\theta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\theta,\delta))\big] \geq 0$.

The next and final result of this section sheds light on the critical role of information about agents' externality assessments.

[30]The latter maximum value exists as $v_i$ and $k^*$ are twice continuously differentiable and $K$ is compact.

**Theorem I.3** *Suppose externality types are common knowledge. Then any differentiable allocation rule* $k^* : \Theta \times \Delta \to K$ *satisfying* $\frac{\partial}{\partial \theta_i} \mathbb{E}_{\theta_{-i}}\big[v_i(k^*(\theta,\delta))\big] \geq 0$ *for all* $(\theta_i, \delta) \in \Theta_i \times \Delta$ *and all* $i \in \{1,2\}$ *can be strongly Bayesian implemented with an ex post budget-balanced mechanism.*

**Proof.** Let $k^* : \Theta \times \Delta \to K$ be an allocation rule satisfying $\frac{\partial}{\partial \theta_i} \mathbb{E}_{\theta_{-i}}\big[v_i(k^*(\theta,\delta))\big] \geq 0$ for all $(\theta_i, \delta)$ and all $i$. For agents $i \in \{1,2\}$ of *commonly known* externality types $\delta = (\delta_1, \delta_2) \in \Delta$, define the function $S_i : \Theta \times \Delta \to \mathbb{R}$ by

$$S_i(\hat{\theta},\delta) = \int_{\theta_i^{\min}}^{\hat{\theta}_i} v_i(k^*(s,\hat{\theta}_{-i},\delta))\, ds - \pi_i(k^*(\hat{\theta},\delta)\,|\,\hat{\theta}_i) - \delta_i \pi_{-i}(k^*(\hat{\theta},\delta)\,|\,\hat{\theta}_{-i}). \tag{I.28}$$

Then the budget-balanced transfer scheme $T^* = (t_1^*, t_2^*) : \Theta \times \Delta \to \mathbb{R}^2$ defined by

$$\begin{aligned} t_1^*(\hat{\theta},\delta) &= \frac{1}{1-\delta_1}\Big[S_1(\hat{\theta},\delta) - \mathbb{E}_{\theta_1}\big[S_1(\theta_1,\hat{\theta}_2,\delta)\big]\Big] \\ &\quad + \frac{1}{1-\delta_2}\Big[-S_2(\hat{\theta},\delta) + \mathbb{E}_{\theta_2}\big[S_2(\hat{\theta}_1,\theta_2,\delta)\big]\Big], \end{aligned} \tag{I.29}$$

$$\begin{aligned} t_2^*(\hat{\theta},\delta) &= \frac{1}{1-\delta_1}\Big[-S_1(\hat{\theta},\delta) + \mathbb{E}_{\theta_1}\big[S_1(\theta_1,\hat{\theta}_2,\delta)\big]\Big] \\ &\quad + \frac{1}{1-\delta_2}\Big[S_2(\hat{\theta},\delta) - \mathbb{E}_{\theta_2}\big[S_2(\hat{\theta}_1,\theta_2,\delta)\big]\Big] \end{aligned} \tag{I.30}$$

strongly Bayesian implements $k^*$. For an extensive proof of this claim as well as a derivation of $T^*$, see the Appendix. ■

As implied by condition (A.16) in the proof of Lemma I.1, the sufficient condition of Theorem I.3 is also necessary.

By Theorem I.3, it is not externality assessments per se that constrain the implementability of allocation rules, but rather the asymmetry of information about them. However, in light of the above quoted 'Wilson doctrine', Theorem I.3 is merely of theoretical relevance.

For the sake of completeness, I should briefly comment on the feasibility of budget-balanced *ex post* implementation. Jehiel et al. (2006) have shown that only constant allocation rules are ex post implementable, irrespective of budget balance. Their model framework entails the one of the present study, the only exception being that, in their model, agents do not internalize the distributive effects of transfers. By the following argument, their result applies nevertheless to the model framework of Section I.3: Consider a social choice rule $(k^*, t_1, t_2)$, consisting of an allocation rule $k^*$ and transfers $(t_1, t_2)$ in the manner of Section I.3. Expand their model framework by allowing for monetary transfers $(t_1', t_2')$ the distributive effects of which are not internalized by agents. By Jehiel et al. (2006), the social choice rule $(k^*, t_1, t_2)$ can be ex post implemented with some transfer scheme $(t_1', t_2')$ only if $(k^*, t_1, t_2)$ is constant, which requires $k^*$ to be constant. This implication holds in particular for the case of $(t_1', t_2') = (0, 0)$. As a sensitive allocation rule $k^*$ is non-constant by definition, the (unfortunate) conclusion is that there exists no sensitive allocation rule that could be ex post implemented. Hence, while budget-balanced Bayesian implementation of sensitive allocation rules can (and must) be externality-robust, the assumption that payoff-type distributions are common knowledge remains critical.

## I.5. Bargaining with Side Payments

This section applies the results obtained above to the following question: How, by what means and what ends, do two agents come to an agreement upon the division of a given 'pie' which is currently owned by neither of them? With 'means' I refer to the bargaining process, with 'ends' to those allocations that are 'feasible' under that process. In particular, how is the feasibility of means and ends restricted when assuming that agents are privately informed about how they value shares of 'pie' and how they assess the externalities, tangible or intangible, that their opponent's share might impose on them?

The bargaining literature can be broadly separated into two strands, one focusing on means, the other one on ends. The 'means'-strand, starting with Rubinstein (1982), starts out from bargaining rules and takes ends as equilibrium outcomes of the respective non-cooperative game.[31] The 'ends'-strand, starting earlier with Nash (1950), is often referred to as 'axiomatic bargaining' and asks for reasonable, axiomatized properties that an al-

[31] See Ausubel, Cramton, and Deneckere (2002) for a survey on non-cooperative bargaining under incomplete information.

location rule, the *bargaining solution*, should possess.[32] Naturally, these properties are preference-contingent, which makes preference revelation a critical issue. Of course, these strands of literature have not been disjoint. For instance, Myerson (1979) has shown that, in a general setting which comprises private information about externality assessments, there exists a unique bargaining solution that is Bayesian incentive-compatible: it maximizes the *generalized Nash product* of Harsanyi and Selten (1972).[33]

In the following, I discuss strategic bargaining from a mechanism design perspective, for "this allows us to identify properties shared by all Bayesian equilibria of any bargaining game" (Ausubel, Cramton, and Deneckere, 2002). I ask which bargaining solutions are strongly Bayesian incentive-compatible if utility is transferable by means of side payments between agents.[34]

Consider two agents, 1 and 2, who bargain over the division of a 'pie' of size 1. Modify the model framework of Section I.3 by assuming that, for all $k \in [0,1]$, $v_1(k) = v(k)$ and $v_2(k) = v(1-k)$, where $v : [0,1] \to [0,1]$ is twice continuously differentiable and satisfies $v(0) = 0$, $v(1) = 1$, $v' > 0$, and $v'' < 0$. Let $h_1(k) = h_2(k) = 0$ for all $k$. From their shares $k$ and $1-k$, respectively, and transfers $t_1$ and $t_2$, agents 1 and 2 draw ex post utilities

$$\begin{aligned} u_1(k) &= \left[\theta_1 v(k) + t_1\right] + \delta_1 \cdot \left[\theta_2 v(1-k) + t_2\right], \\ u_2(k) &= \left[\theta_2 v(1-k) + t_2\right] + \delta_2 \cdot \left[\theta_1 v(k) + t_1\right]. \end{aligned}$$

By Theorem I.1, the only *sensitive* sharing rule, or bargaining solution, that can be *strongly* Bayesian implemented through budget-balanced transfers is the one associated with externality-ignoring utilitarianism: $k^*(\theta) = \arg\max_{k \in [0,1]} \theta_1 v(k) + \theta_2 v(1-k)$. The

[32] See Thomson (1994) for a survey.

[33] Harsanyi and Selten (1972) propose maximization of the *generalized Nash product* as an axiomatic solution to bargaining under incomplete information. Notice that the *generalized Nash product* takes type distributions explicitly into account. The welfare judgment it entails thus depends on what bargainers' types *could have been* and not merely on what agents' types *are ex post*. Consequently, the result of Myerson (1979) hinges on the assumption that type distributions are common knowledge—an assumption in conflict with the 'Wilson doctrine'. As will be shown below, bargaining procedures can at least be externality-robust if one allows for side payments.

[34] The results are also informative for "pure" bargaining (i.e., if utility is not transferable), since side payments can be zero if the bargaining solution is incentive-compatible on its own.

respective transfer scheme is necessarily of *AGV-type*: If agents 1 and 2 claim to be of types $(\hat{\theta}_1, \hat{\delta}_1)$ and $(\hat{\theta}_2, \hat{\delta}_2)$, transfers are given by

$$\begin{aligned} t_1(\hat{\theta}, \hat{\delta}) &= \mathbb{E}_{\theta_2}\big[\theta_2 v(1 - k^*(\hat{\theta}_1, \theta_2))\big] - \mathbb{E}_{\theta_1}\big[\theta_1 v(k^*(\theta_1, \hat{\theta}_2))\big] + s(\hat{\theta}, \hat{\delta}), \\ t_2(\hat{\theta}, \hat{\delta}) &= \mathbb{E}_{\theta_1}\big[\theta_1 v(k^*(\theta_1, \hat{\theta}_2))\big] - \mathbb{E}_{\theta_2}\big[\theta_2 v(1 - k^*(\hat{\theta}_1, \theta_2))\big] - s(\hat{\theta}, \hat{\delta}), \end{aligned}$$

where $s$ must be chosen such that $\mathbb{E}_{\theta_{-i}, \delta_{-i}}[s(\theta, \delta)]$ is constant on $\Theta_i \times \Delta_i$ for each $i$, such that externality assessments are left strategically inoperative. That is, negotiations must focus on private payoffs, irrespective of externalities. When letting $s = 0$, as bargainers' externality assessments might not be common knowledge, the transfer scheme indicates that agents make mutual concessions which amount to the expected externalities they impose on one another under the sharing rule $k^*$.

The necessity of externality robustness seems particularly plausible in the range of conflict resolution. An arbitrator, seeking to resolve dispute between hostile parties, should rather claim "Let's focus on the issue!" than care about who likes or dislikes whom how much (and is thus more or less altruistic or spiteful).

The results of Sections I.3 and I.4 preclude the most prominent solutions to axiomatic bargaining from being *strongly* Bayesian implemented without incentive costs; if at all, they are Bayesian implementable through budget-balanced transfers only for very specific type distributions.

**Proposition I.5** *The bargaining solutions of Nash (1950), Kalai (1977), and Kalai and Smorodinsky (1975), all of these either externality-sensitive or externality-ignoring, cannot be strongly Bayesian implemented through budget-balanced transfers. The opposite would hold if externality assessments were common knowledge.*[35]

**Proof.** Notice first that condition (I.7) of Lemma I.1 implies that a partially differentiable bargaining solution $k^* : \Theta \times \Delta \to [0, 1]$ which *does not* maximize the sum of private payoffs

[35]These bargaining solutions are well-defined when presuming $\Delta_i \subset [0, 1]$ for Nash and $\Delta_i \subset [-1, \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}]$ for Kalai; for Kalai-Smorodinsky, the first part of the Proposition presumes $\Delta_i \subset [-1, \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}]$, whereas the second part presumes $\Delta_i \subset [0, \frac{\theta_i^{\min}}{2\theta_{-i}^{\max}}]$.

is strongly Bayesian implementable in a budget-balanced way *only if* the following holds for all $(\theta, \delta)$:[36]

$$\operatorname{sgn}\left(\frac{\partial k^*}{\partial \theta_1}\frac{\partial k^*}{\partial \theta_2}\right) = -\operatorname{sgn}\left(\frac{\partial k^*}{\partial \delta_1}\frac{\partial k^*}{\partial \delta_2}\right). \tag{I.31}$$

The externality-sensitive *Nash solution* is given by

$$k^*(\theta, \delta) = \arg\max_{k\in[0,1]} \left[\theta_1 v(k) + \delta_1\theta_2 v(1-k)\right]\cdot\left[\theta_2 v(1-k) + \delta_2\theta_1 v(k)\right]. \tag{I.32}$$

The externality-sensitive *Kalai solution*, in the manner of Rawls (1971), requires to maximize the minimum of agents' ex post utilities. Consider its externality-sensitive egalitarian version: $k^* = k^*(\theta, \delta)$ such that $u_1(k^*) = u_2(k^*)$. This is equivalent to $k^*$ satisfying[37]

$$0 = (\theta_1 - \delta_1\theta_2)v(1-k^*) - (\theta_2 - \delta_2\theta_1)v(k^*) = F(k^*, \theta, \delta). \tag{I.33}$$

The externality-sensitive *Kalai-Smorodinsky solution* requires $k^*$ to equalize the ratio of agents' ex post utilities and the ratio of agents' maximum potential gains: $\frac{u_1(k^*)}{u_2(k^*)} = \frac{u_1(1)}{u_2(0)}$. This is equivalent to $k^*$ satisfying[38]

$$0 = \theta_2(\theta_1 - \delta_1\theta_2)v(1-k^*) - \theta_1(\theta_2 - \delta_2\theta_1)v(k^*) = G(k^*, \theta, \delta). \tag{I.34}$$

The respective externality-ignoring versions of (I.32) to (I.34) are obtained when letting $\delta_1 = \delta_2 = 0$ in each of them. These externality-ignoring bargaining solutions violate condition (I.31), since then $\frac{\partial k^*}{\partial \delta_i} = 0$, whereas $\frac{\partial k^*}{\partial \theta_i} \neq 0$.

Assuming non-negative externality types, the Nash solution is not strongly Bayesian implementable through budget-balanced transfers, due to Proposition I.1(iv) and Theorem I.1, whereas the second part of the Proposition is implied by Theorem I.3.

[36]Specifically, multiplying (I.7) for $i = 1$ with (I.7) for $i = 2$ and then applying the sign function to both sides of the resulting identity yields the condition: $\operatorname{sgn}\left(\frac{\partial k^*}{\partial \theta_1}\frac{\partial k^*}{\partial \theta_2}\right) = \operatorname{sgn}\left(\frac{\partial v_1(k^*)}{\partial \delta_1}\frac{\partial v_2(k^*)}{\partial \delta_2}\right)$. In the present context, this condition is equivalent to (I.31).

[37]Condition (I.33) is well-defined on $\Theta \times \Delta$ if and only if $\delta_i^{\max} < \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}$ for all $i$: On the one hand, (I.31) has a solution $k^*$ if and only if $(\theta_1 - \delta_1\theta_2), (\theta_2 - \delta_2\theta_1) > 0$ or $(\theta_1 - \delta_1\theta_2), (\theta_2 - \delta_2\theta_1) < 0$; however, the latter inequality would imply that $(1 - \delta_1\delta_2)\theta_i < 0$, which contradicts the assumptions on $\Theta \times \Delta$.

[38]As before, condition (I.34) is well-defined on $\Theta \times \Delta$ if and only if $\delta_i^{\max} < \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}$ for all $i$.

It will be made clear in the Appendix that the bargaining solutions (I.33) and (I.34) both satisfy the following conditions: $\text{sgn}(\frac{\partial k^*}{\partial \delta_1}\frac{\partial k^*}{\partial \delta_2}) = -1 = \text{sgn}(\frac{\partial k^*}{\partial \theta_1}\frac{\partial k^*}{\partial \theta_2})$ for all $(\theta, \delta)$ with $\delta_i < \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}$, which violates condition (I.31). Furthermore, $\frac{\partial k^*}{\partial \theta_2} < 0 < \frac{\partial k^*}{\partial \theta_1}$ if $\Delta_i \subset [-1, \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}]$ in case of (I.33) and $\Delta_i \subset [0, \frac{\theta_i^{\min}}{2\theta_{-i}^{\max}}]$ in case of (I.34). The latter condition implies in particular that $\frac{\partial v_i(k^*)}{\partial \theta_i} > 0$ for all $i$, such that $\frac{\partial}{\partial \theta_i}\mathbb{E}_{\theta_{-i}}\big[v_i(k^*(\theta,\delta))\big] > 0$; hence, Theorem I.3 gives proof of the second part of the Proposition. ■

## I.6. Holistic Social Welfare Measures

Up to this point, I have restricted attention to the social welfare judgment inherent to the allocation rule. To which extent does the result of Theorem I.1 expand to social welfare judgments that are *holistic* in the sense that they incorporate the distributive effects of a transfer scheme? With Theorem I.1 at hand, it is easy to answer this question.

Consider a differentiable function $W : \mathbb{R}^4 \to \mathbb{R}$ and define $V : K \times \mathbb{R}^2 \to \mathbb{R}$ by

$$V(k, t_1, t_2) = W\big(\pi_1(k, t_1 \,|\, \theta_1)\,,\, \delta_1\pi_2(k, t_2 \,|\, \theta_2)\,,\, \pi_2(k, t_2 \,|\, \theta_2)\,,\, \delta_2\pi_1(k, t_1 \,|\, \theta_1)\big),$$

where $\pi_i(k, t_i \,|\, \theta_i) = \theta_i v_i(k) + h_i(k) + t_i$. Suppose $W$ is an *ex post social welfare measure* in that it is invariant to changes in type distributions. Assume also that $W$ satisfies

$$\frac{\partial W}{\partial \pi_1} = [W_1 + \delta_2 W_4] > 0 \quad \text{and} \quad \frac{\partial W}{\partial \pi_2} = [W_3 + \delta_1 W_2] > 0. \tag{I.35}$$

The social choice rule $(k^*, t_1^*, t_2^*)$, with allocation rule $k^* : \Theta \times \Delta \to K$ and transfer scheme $(t_1^*, t_2^*) : \Theta \times \Delta \to \mathbb{R}^2$, is budget-balanced and maximizes $V$ if and only if $t_2^* = -t_1^*$ and

$$(k^*, t_1^*) = \arg\max_{(k,t_1)\in K\times\mathbb{R}} V(k, t_1, -t_1). \tag{I.36}$$

Assuming $W$ allows for interior solutions, $(k^*, t_1^*)$ satisfies the first-order conditions

$$0 = \frac{\partial V(k^*, t_1^*, -t_1^*)}{\partial k} = [W_1 + \delta_2 W_4]\frac{d\pi_1(k^* \,|\, \theta_1)}{dk} + [W_3 + \delta_1 W_2]\frac{d\pi_2(k^* \,|\, \theta_2)}{dk}, \tag{I.37}$$

$$0 = \frac{\partial V(k^*, t_1^*, -t_1^*)}{\partial t_1} = [W_1 + \delta_2 W_4] - [W_3 + \delta_1 W_2], \tag{I.38}$$

where $\pi_1(k^* \mid \theta_1) = \theta_i v_i(k) + h_i(k)$. Conditions (I.35), (I.37), and (I.38) jointly imply that the socially efficient allocation rule is necessarily consistent with externality-ignoring utilitarianism: $k^*(\theta, \delta) = \arg\max_{k \in K} \pi_1(k \mid \theta_1) + \pi_2(k \mid \theta_2)$. In other words, under holistic social welfare measures, social choice differs merely in the extent of redistributive taxation. The problem thus reduces to the question: Which social welfare judgments yield redistributive tax tariffs that are Bayesian incentive-compatible?

By Theorem I.1, the optimal allocation rule $k^*$ can be Bayesian implemented in a budget-balanced way if and only if transfers are of AGV-type. As AGV-type transfers vary with changes in the distribution of types, the *ex post social welfare measure* $W$ must be invariant to changes in transfers. In other words, agents' private payoffs must be perfect substitutes from a social planner's point of view. This proves the following theorem.

**Theorem I.4** *A budget-balanced social choice rule, $(k^*, t^*, -t^*)$, that is interior solution to the maximization of a differentiable ex post social welfare measure $W$ satisfying condition (I.35) is Bayesian incentive-compatible if and only if $W$ is consistent with externality-ignoring utilitarianism. The respective mechanism is of AGV-type.*

Theorem I.4 applies in particular to the welfare measures listed in Proposition I.1.

A final remark can be made on Rawlsian justice (Rawls, 1971). While the non-differentiable (and non-sensitive) Rawlsian maximin welfare function does not meet with the presumptions of the above analyses, Theorem I.1 still proves useful to obtain the following result.

**Proposition I.6** *A budget-balanced social choice rule, $(k^*, t^*, -t^*)$, satisfying Rawls' maximin principle, inclusive or exclusive of externalities, is not Bayesian incentive-compatible.*

**Proof.** Consider the maximin principle inclusive of externalities and let

$$(k^*, t^*) = \arg \max_{(k,t) \in K \times \mathbb{R}} \min\{\pi_1(k, t \mid \theta_1) + \delta_1 \pi_2(k, -t \mid \theta_2)\, ;\, \pi_2(k, -t \mid \theta_2) + \delta_2 \pi_1(k, t \mid \theta_1)\}.$$

As individual utility is affine in transfers, $t^*$ must equalize utilities:

$$\pi_1(k^* \mid \theta_1) + \delta_1 \pi_2(k^* \mid \theta_2) + (1 - \delta_1) t^* = \pi_2(k^* \mid \theta_2) + \delta_2 \pi_1(k^* \mid \theta_1) - (1 - \delta_2) t^*,$$

where $\pi_1(k^* \,|\, \theta_1) = \theta_i v_i(k) + h_i(k)$. Therefore, $t^* = \frac{1-\delta_1}{2-\delta_1-\delta_2}\pi_2(k^* \,|\, \theta_2) - \frac{1-\delta_2}{2-\delta_1-\delta_2}\pi_1(k^* \,|\, \theta_1)$ and utilities are given by

$$u_1 = u_2 = \frac{1-\delta_1\delta_2}{2-\delta_1-\delta_2}\left[\pi_1(k^* \,|\, \theta_1) + \pi_2(k^* \,|\, \theta_2)\right].$$

Hence, $k^* = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$, since $\delta_i \in (-1, 1)$. By Theorem I.1, transfers must be of AGV-type so as to Bayesian implement $k^*$. As $t^*$ is not of AGV-type, $(k^*, t^*, -t^*)$ is not Bayesian incentive-compatible.

When letting $\delta_i = 0$ in the above line of reasoning, the proof is obtained for the maximin principle exclusive of externalities. ■

## I.7. Conclusion

How agents assess the (in-)tangible externalities that others might impose on them can strongly influence strategic interaction. I have explored ex post Pareto-efficient Bayesian implementation for agents whose externality assessments and private payoffs, exclusive of externalities, are all subject to asymmetric information. Under reasonable assumptions, ex post Pareto-efficient allocations are Bayesian implementable with a budget-balanced mechanism if and only if the social welfare judgment underlying the choice of allocations is that of externality-ignoring utilitarianism. This restriction is caused by the asymmetry of information about agents' externality assessments, as common knowledge of externality assessments allows for budget-balanced Bayesian implementation of (nearly) any allocation rule.

The ex post Pareto-efficient, budget-balanced mechanism corresponding to externality-ignoring utilitarianism necessarily takes the form of the renowned 'expected externality mechanism' due to Arrow (1979) and d'Aspremont and Gérard-Varet (1979). This mechanism is externality-robust in that it leaves externality assessments strategically inoperative. Externality robustness turns out to be not just a desirable property in order to avoid unrealistic common knowledge assumptions, as urged by Wilson (1987): externality robustness is rather necessary from an incentive compatibility point of view. Otherwise, agents would internalize the distributive effects of the mechanism itself, and counterbalancing the associated adverse incentives would come at costs, embodied in the violation of budget balance. As they result from the welfare judgment inherent to an allocation rule, I have called these the incentive costs of welfare judgments.

In the range of conflict resolution, the central result provides a rationale for the common-sense approach most people would adopt when arbitrating between conflicting parties: namely, to *not* condition the arbitration process or the final resolution on the extent to which the opponents despise each other, but to rather "focus on the issue" and to base arbitration merely on how it would affect the opponents' *material* wealth: One may think of how judges approach the resolution of divorce battles, how a mother tends to resolve animosity between her children, or how third-party diplomats try to conciliate rival tribes or nations. The central result implies in particular that, even when allowing for side payments, the most prominent bargaining solutions, namely those of Nash (1950), Kalai (1977), and Kalai and Smorodinsky (1975), are not Bayesian incentive-compatible.

From a more general perspective, the result suggests that public economic policies dedicated to maximize a social welfare measure inconsistent with externality-ignoring utilitarianism do either provide people with adverse incentives (e.g., to reduce their labor supply) or are not budget-balanced, leading either to a waste of money or an increase in public debt.

# Chapter II.

# Friends and Foes at Work: Assigning Teams in a Social Network

## II.1. Introduction

Management consultants, fishery operators, and teachers, too, frequently face a similar task: the division of a group of people into teams. Management consultancies typically serve several clients at once, advising each with a different team of experts. Fishery operators commonly send out several boats at once, in hope of finding the most fruitful fishing grounds. When open house day is near, teachers utilize teamwork in arts classes in order to obtain multiple exhibits that impress parents and, hopefully, donors. Parallel production of this manner is widely used. Often, however, a principal can only observe the performance of a team as a whole. Under such conditions, agents cannot be held accountable for their individual contributions to a team's success or failure; they are presented with the temptation to free-ride on the contributions of their fellow team members.

As is typical for the ubiquitous *project teams*, teamwork is often short-lived, while the group of agents as a whole persists. In persistent groups, however, people develop interpersonal relationships. Their relationships can affect their willingness to cooperate within a team and thereby help to either diminish or increase their incentives to free-ride. Taking an outside perspective, these interpersonal relationships constitute a social network. When dividing the group into teams, this social network will affect both the productivity of teams as well as the agents' subjective well-being.

How then should a principal assign agents to teams when she seeks maximum overall productivity, or maximum profit? And, considering she has no information on the agents' social network, can she leave the decision of who teams up with whom to the agents? Or should she acquire the relevant information and take action? The aim of this study is to answer both questions: Given the agents' social network of interpersonal relationships, what is the efficient assignment of teams? And is there a universal mechanism guaranteeing efficiency while delegating responsibility for team assignment to the agents?

A simple narrative helps to illustrate the economic problem involved. The operator of a fishery is equipped with two boats and four fishermen: John, Joe, Jim, and Jimmy. The boats each have to be run by a crew of two. At the end of a working day, the operator, who herself stays in the harbor, observes every crew's catch and pays each crew a share of the respective market return. Crew members each receive half of their crew's pay, since neither of them can plausibly convey having had a larger impact on their crew's catch than the other crew member. As always, Jim teams up with his best friend Jimmy, and John teams up with his best friend Joe. But for some reason unrelated to work, conflict breaks out between John and Joe. The formerly altruistic attitudes that John and Joe had toward one another turn into spite. The operator now faces a tradeoff, in case she knows of this conflict: Having John and Joe quarrel all day instead of hauling in nets would decrease productivity. Reassigning crews to consist of John and Jim, and Joe and Jimmy, might prevent the loss of productivity from John and Joe working together. However, such an intervention would come at a loss: separation of the perfectly cooperating Jim and Jimmy. Whether or not reassignment increases overall productivity depends on the fishermen's willingness to cooperate within the alternative crews. Given these alternatives, the question is whether the highly motivated crew would compensate for the poorly motivated crew, or whether the conflicting parties should be separated. The answer to this question must hinge on the determinants of production; it depends on how exactly an improvement of the interpersonal relationship within a crew translates into higher productivity of said crew. When staffing the crews, a fully informed operator accounts for both the determinants of production and the social network of her staff.

Constantly being informed on the interpersonal relationships in her staff is costly for the operator. She might want to avoid those costs by simply leaving the assignment of crews to the fishermen themselves. In addition, such delegation of responsibility might avoid the hidden costs of control (Falk and Kosfeld, 2006) and increase the fishermen's motivation (Charness et al., 2012).[1] Would the fishermen self-select into the most profitable composition of crews? To be more precise, is there a mechanism that incentivizes them to do so? - Suppose the perfectly cooperating Jim and Jimmy could indeed compensate for the crew of conflicting John and Joe, as compared to the feasible alternatives. The operator would thus prefer not to change crew composition. However, it is intuitive to think of the fishermen's social network as such that every individual prefers the separation of the conflicting parties: John and Joe may prefer to be separated for their own sakes; Jim and Jimmy may be willing to sacrifice their own success in order to support their fellow fishermen. Could delegation succeed if the fishermen's unanimous preferences are in opposition to those of the operator?

Early theoretical literature on teamwork focused on moral hazard in the context of purely self-interested agents.[2] As observed in experiments on private contributions to public goods (the laboratory analog to real world teamwork), a substantial share of individuals exhibit social preferences that can be related to concepts of altruism and spite (Saijo and Nakamura, 1995, Levine, 1998, and Andreoni and Miller, 2002).[3] In a field experiment with fishermen, Carpenter and Seki (2011) demonstrate that the social preferences observed in the laboratory are positively correlated to individual efforts in real world

[1] Flores-Fillol, Iranzo, and Mane (2017) find similar results in the range of teamwork. Other studies emphasize the drawbacks resulting from delegation: While abstracting from any form of social concerns, or intrinsic motivation, Bester and Krähmer (2008) argue with the help of a principal-single agent model that delegation, generally, cannot lead to efficient outcomes. Ingvaldsen and Rolfsen (2012) show empirically that delegation in large organizations can lead to inefficiencies associated with between-team coordination.

[2] These studies investigate which incentive schemes help to overcome moral hazard in teams, and under which conditions team incentives implement first-best or second-best production levels (e.g., Holmstrom, 1982, Itoh, 1991, and McAfee and McMillan, 1991).

[3] These preferences tend to be conditional on other participants' willingness to cooperate (Fischbacher, Gächter, and Fehr, 2001, and Van Dijk, Sonnemans, and van Winden, 2002) and involve the propensity to punish free-riders in order to obtain socially optimal outcomes (e.g., Fehr and Gächter, 2002).

teamwork environments.[4] Babcock et al. (2015) observe that team incentives, when combined with the opportunity for regular social interaction, induce participants to be more committed than under individual-based incentives. Since production complementarities are absent, the authors attribute this finding to social preferences in the broader sense. Bandiera, Barankay, and Rasul (2010) find that the network of workplace friendships affects the performance of workers: even when pay is based on individual performance, and production complementarities are absent, workers are on average more productive when those they are socially tied to are close to them during work. Mas and Moretti (2009), in a field experiment with supermarket cashiers, identify positive spillover effects from highly productive workers to those who are less so; they attribute this finding to social pressure through mutual monitoring among those colleagues who frequently work during the same shift.[5]

These field studies provide strong evidence that the interpersonal relationships in the workplace should be accounted for when staffing crews, project teams, and shifts.[6] However, the staffing of teams and the mode in which it affects an organization's performance at large cannot exclusively be evaluated within the context of social networks. Teamwork and the division of staff into teams are constrained by the organizational context in which they take place. This involves the organization of production on the team level, the difficulty of production on the individual level, and the incentives provided by the principal.

I assume the agents' social network is determined by their altruistic or spiteful interpersonal preferences. These preferences can range from strong spite to strong altruism; they are assumed to be mutual between every two agents, are exogenously given in the moment of team assignment, and remain unaffected by it. Considering piece rate compensation

[4] In particular, they find that altruistic attitudes evolve among people who are frequently exposed to team incentives in their everyday work, providing evidence for the theory of Rotemberg (1994).

[5] Their study thus provides evidence for the respective theoretical prediction by Kandel and Lazear (1992).

[6] Other studies on the role and adequate 'use' of social networks in organizations focus on communication ties and the efficiency of information flow (Oh, Chung, and Labianca, 2004, and Balkundi and Harrison, 2006), on the distribution of skills within and across teams (Hamilton, Nickerson, and Owan, 2003), or on the composition of teams regarding the demographic or psychological characteristics of team members (Neuman and Wright, 1999, Rulke and Galaskiewicz, 2000, Reagans, Zuckerman, and McEvily, 2004, and Elfenbein and O'Reilly, 2007).

for teams and within-team efforts that are perfectly substitutable, I identify rules for efficient team assignment.[7] These rules vary with the structure of individual effort costs. For convex (concave) marginal costs of effort, efficient team assignment follows a maximin (maximax) rule with regard to the agents' willingness to cooperate. One practical implication of this finding is that team assignment should focus on the separation of conflicting parties if the production of any additional unit of output is very costly.

Utilizing these rules, I discuss the efficiency of delegation. The self-selection of agents into teams, according to a certain mechanism, imposes an externality on the principal. While the principal wants to extract a rent from team assignment, the agents are not following their material self-interest alone, they also take into account the effects of team assignment on their colleagues' wealth.[8] I consider mechanisms of *delegation* by which I refer to mechanisms that involve the agents in a strategic game the rules of which exclusively depend on the agents' preference orderings over team composition (for instance, majority voting). Delegation mechanisms, by this requisite, are income neutral: their application does not affect the agents' wealth beyond the implementation of a specific team assignment.

I show that a universal mechanism guaranteeing efficiency while delegating responsibility for team assignment to the agents does not exist. In fact, there exist social networks in which the agents unanimously prefer the implementation of a team assignment that does

[7] An alternative form of incentives for teams has been proposed by Gershkov, Li, and Schweinzer (2009). They show that moral hazard in teams can be overcome through a Tullock contest between team members; if the signals on the agents' individual efforts are not perfectly correlated, first-best efficient efforts can be implemented. Notice that relative performance incentives impose a negative externality of every agent's effort on the other agent's pay, reversing the effects of within-team altruism and spite as compared to piece rate compensation for teams. On the other hand, if the principal observes spite between certain agents, she might even want to exploit this spite by imposing a contest between appropriate teams. See Bandiera, Barankay, and Rasul (2013) for a field experiment on how tournaments between teams affect team performance and overall productivity. However, other empirical studies suggest that contests might increase or even create spite between workers, potentially leading to sabotage (e.g., Goette et al., 2012, and Charness, Masclet, and Villeval, 2013). When imposing contests within or between teams, the questions addressed in this chapter arise similarly.

[8] Due to this externality, the analysis of endogenous team formation in the workplace goes beyond theories on the endogenous formation of networks and network stability (Dutta, Ghosal, and Ray, 2005, and Page, Wooders, and Kamat, 2005), on networks of endogenous externalities among agents (Bramoullé and Kranton, 2007), and on equilibrium behavior in network games in general (Jackson and Wolinsky, 1996, and Galeotti et al., 2010). This chapter adds a principal-agent perspective to social network theory.

not maximize overall productivity or the principal's profit. Such a Pareto dominant team assignment from the agents' point of view would inevitably be the outcome of a Nash equilibrium under any kind of delegation mechanism. Hence, there exists no delegation mechanism for which the outcome of *every* Nash equilibrium is the efficient assignment of teams. The respective examples share the trait that the social network is either conflict-laden or, in the absence of spite, asymmetric (that is, at least one team assignment results in one highly cooperative and one poorly cooperative team). Interestingly, even if all the agents are altruistic toward one another, they might unanimously opt for a mode of production that is inefficient overall.

Another way of overcoming the acquisition of information is the *pooling of incentives* by paying every agent an equal share of the market return on the overall output. Team assignment could be arbitrary in this case. Intuitively, pooling would increase the agents' incentives to free-ride. One would expect that there is at least one composition of teams, with teams of two agents being paid according to their own team output, that is more efficient than pooling. Counterintuitively, there do exist social networks for which the pooling of incentives can be efficient. These social networks are all characterized by the presence of a 'spiteful outcast', an agent who is the recipient of all the other agents' spite and who reciprocates this spite. Nevertheless, knowing whether or not pooling is the best response to the agents' social network requires the principal to be informed.

From a purely contract theoretical point of view, team assignment can be arbitrary with respect to the agents' social network. If the principal knows what team output to expect in the case of efficiently working agents, then efficient effort levels can be enforced through Holmstrom's (1982) budget breaking rule: If team output is as large as if every team member had produced efficiently, then team members are paid according to their reservation utilities; otherwise, they receive no pay. Under such a regime, exerting the efficient efforts constitutes a Nash equilibrium regardless of the team members' interpersonal preferences (except for unrealistically strong spite between agents). Yet often, a principal does not know what the outcome of a team's work could have been at best: Having returned to the harbor with a poor catch, foes John and Joe can simply claim that they were unlucky, the big shoal must have been somewhere else. The operator has no means of falsifying their lie. Consequently, real world fishermen are often paid a share of the market return on their catch. Similarly, the teacher hoping for high quality pieces of art to be exhibited on open house day just might not know beforehand which pieces can be rated 'high quality'. In this respect, understanding team assignment in social networks is first and foremost a matter of practical relevance.

## II.2. The Model

A principal faces a group of four agents, $\{a, b, c, d\}$. Production in the principal's firm requires collaboration in teams of two. The group of agents has to be subdivided into these teams. For a combination of agents, $\{i, j, k, l\} = \{a, b, c, d\}$, denote by $[(ij)(kl)]$ the formation of two teams, one containing $i$ and $j$, the other one $k$ and $l$. Exactly three of such *team assignments* are feasible, and one of them must be implemented:

$$[(ij)(kl)] \in \{[(ab)(cd)], [(ac)(bd)], [(ad)(bc)]\}. \tag{II.1}$$

Once a team $(ij)$ is formed, $i$ and $j$ make simultaneous effort choices $x_i, x_j \in [0, \infty)$. The function of individual effort costs, $C : [0, \infty) \to [0, \infty)$, is the same for all the agents. It is sufficiently often differentiable and satisfies $C(0) = C_x(0) = 0$; $C_x, C_{xx} > 0$; and $\lim_{x \to \infty} C_x(x) = \infty$.

The principal can only observe team output. Within teams, individual efforts are perfectly substitutable. Team effort $x_i + x_j$ transforms directly into team output, which the principal sells at a market price of one per unit. Teams receive piece rate compensation for their own team output, of which each team member receives an equal share. The piece rate is exogenously given and the same for both teams. With piece rate $w \in (0, 1]$, team $(ij)$ receives compensation $w \cdot (x_i + x_j)$ for its team output. Consequently, agent $i$ ends up with material wealth

$$\pi_i = \frac{1}{2} w(x_i + x_j) - C(x_i). \tag{II.2}$$

The principal ends up with a profit of $\Pi_{(ij)} = (1 - w)(x_i + x_j)$ from team $(ij)$ and a profit of $\Pi_{(kl)} = (1 - w)(x_k + x_l)$ from team $(kl)$.

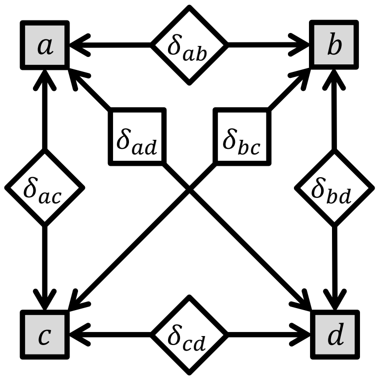


Figure II.1.: *The social network of interpersonal preferences, $\delta_{ij} \in [-1, 1]$, in the group of agents $\{a, b, c, d\}$.*

Each agent $i$ maximizes utility

$$u_i = \sum_{j \in \{a,b,c,d\}} \delta_{ij} \pi_j, \tag{II.3}$$

with $\delta_{ii} = 1$ and $\delta_{ij} \in [-1, 1]$ for all $j \neq i$. That is, every agent has a normalized valuation of their own wealth and has altruistic or spiteful valuations of their coworkers' wealth.[9] I refer to $\delta_{ij}$ as the *degree of altruism* between $i$ and $j$. The degrees of altruism between every two agents are assumed to be mutual, $\delta_{ij} = \delta_{ji}$ for all $i, j$.[10] They are exogenously given, will be unaffected by team assignment, and are common knowledge among agents. They determine an interpersonal structure that I refer to as the *social network* $\{a, b, c, d\}$. Figure II.1 illustrates the so defined social network.

Suppose team assignment $[(ij)(kl)]$ has been implemented. Then agent $i$ maximizes utility $u_i = \pi_i + \delta_{ij}\pi_j + \delta_{ik}\pi_k + \delta_{il}\pi_l$ with respect to individual effort. By (II.2), $i$'s marginal utility of individual effort $x_i$ is

$$\frac{du_i}{dx_i} = \left(\frac{1}{2} + \frac{1}{2}\delta_{ij}\right) w - C_x(x_i), \tag{II.4}$$

[9] A similar approach is taken by Brunner and Sandner (2012) who investigate under which constellations of the degrees of altruism a principal's profit is maximal; in their model, moral hazard in teams is absent, and the principal is fully informed.

[10] This mutuality can be interpreted as the result of reciprocity between people who interact frequently. For a survey on reciprocal behavior see Fehr and Schmidt (2006).

with marginal return on effort $\left(\frac{1}{2}+\frac{1}{2}\delta_{ij}\right)w_{ij}$ and marginal costs of effort $C_x(x_i)$. In team assignment $[(ij)(kl)]$, $i$'s effort choice leaves $k$'s and $l$'s material wealth unaffected, and vice versa. Agent $i$'s effort choice is thus independent of his social preferences toward the members of the other team. The marginal return on effort reflects the fact that only half of $i$'s effort transforms into effective return for $i$, the other half benefits co-worker $j$. The weight $\frac{1}{2}+\frac{1}{2}\delta_{ij}$ measures $i$'s valuation of this externality. I refer to

$$m_{ij}=\frac{1}{2}+\frac{1}{2}\delta_{ij} \tag{II.5}$$

as *$i$'s motivation to cooperate with $j$*; in brief, *$i$'s motivation.* The stronger $i$'s altruism towards $j$ is the greater is $i$'s motivation to cooperate with $j$.[11] Marginal return on effort, $m_{ij}w$, depends on both the social incentive to engage in production, $m_{ij}$, and the material incentive to do so, $w$. By mutuality, $m_{ij}=m_{ji}$. Agent $i$ maximizes utility by exerting an effort of

$$x_i^*=C_x^{-1}(m_{ij}w)\in[0,\infty). \tag{II.6}$$

Inverse marginal costs of effort, $C_x^{-1}$, are strictly increasing in marginal return and, therefore, strictly increasing in the motivation to cooperate. The stronger $i$'s altruism towards $j$ is the more will $i$ produce: $dx_i^*/d\delta_{ij}>0$. Notice that perfect spite between team members, $\delta_{ij}=-1$, implies zero team output.

Team assignment $[(ij)(kl)]$ leaves the principal a profit of

$$\Pi_{[(ij)(kl)]}=2(1-w)\left[C_x^{-1}(m_{ij}w)+C_x^{-1}(m_{kl}w)\right]. \tag{II.7}$$

The principal's objective is to maximize this profit with respect to the feasible team assignments $[(ab)(cd)]$ , $[(ac)(bd)]$, and $[(ad)(bc)]$. I assume that every agent has no other option but to agree upon the principal's team assignment decision. I make this

[11] An alternative reading of the parameters $m_{ij}$ is that they do not represent the agents' willingness to cooperate but rather their capability of doing so. This capability might depend on team familiarity (Huckman, Staats, and Upton, 2009), mutual trust (Moldoveanu and Baum, 2011), or personality traits (Neuman and Wright, 1999). Analytically, the task of efficient team assignment remains the same. However, interpersonal preferences are particularly important when it comes to 'pooling the incentives' (Section II.4) and the efficiency of delegation (Section II.5).

assumption to especially account for the pervasive short-lived *project teams* that come together and come apart on a monthly or weekly basis. In the case of fishermen John, Joe, Jim, and Jimmy, crews can be staffed even on a daily basis; the 'outside option' when disagreeing with the operator's staffing decision would then be to get fired and to search for a new job. This might be too large a setback to not to follow the operator's command.

The shape of the principal's profit function (II.7) is the crucial factor for efficient team assignment. It is determined by the shape of $C$. Denote by $\varepsilon_{x_i^*,m_{ij}}$ the elasticity of individual effort with respect to the motivation to cooperate: $\varepsilon_{x_i^*,m_{ij}} = (dx_i^*/x_i^*)/(dm_{ij}/m_{ij})$. Since team members are equally motivated by mutuality, the motivation elasticity of individual effort equals the motivation elasticity of team output, $\varepsilon_{x_i^*,m_{ij}} = \varepsilon_{x_i^*+x_j^*,m_{ij}}$. The principal's profit, the effort cost function, and the motivation elasticity of team output do relate as follows.

**Lemma II.1** *The principal's profit* $\Pi_{[(ij)(kl)]}$ *is strictly increasing in each team's motivation to cooperate,* $m_{ij}$ *and* $m_{kl}$*. With concave (convex) marginal costs of effort,* $C_{xxx} < 0$ *(*$C_{xxx} > 0$*), the principal's profit is convex (concave) in the vector* $(m_{ij}, m_{kl})$ *of motivation per team. With concave (convex) marginal costs of effort, team output is elastic (inelastic) in a team's motivation to cooperate,* $\varepsilon_{x_i^*,m_{ij}} > 1$ *(*$\varepsilon_{x_i^*,m_{ij}} < 1$*).*

**Proof.** The proof is straight forward and omitted therefore. ■

The extent to which a highly motivated team can compensate for an unmotivated team thus depends on the steepness of the effort cost function. The steeper the effort cost function is, the smaller would be a team's additional output in response to a marginal increase in that team's motivation. I impose some regularity on the effort cost function: Either $C_{xxx} < 0$, or $C_{xxx} > 0$. Consequently, team output reacts either elastically or inelastically to changes in a team's motivation.

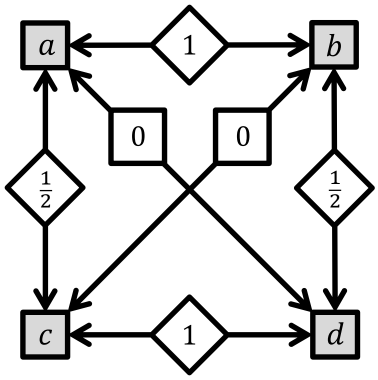


Figure II.2.: *A social network that allows for two highly motivated teams,* $(ab)$ *and* $(cd)$.

## II.3. Principles of Efficient Team Assignment

In this section, I discuss how a principal who has full information on the agents' social network will assign teams in order to realize maximum profit. Notice that, with an exogenously given piece rate for both teams, profit maximization is equivalent to the maximization of overall output and, therefore, to the maximization of production efficiency. I refer to a team assignment as *efficient* if it most profitable among the feasible assignments.

As is obvious from (II.7) and Lemma II.1, the principal prefers a team assignment that allows for a high motivation in both teams.

**Proposition II.1** *If the least motivated team in assignment* $[(ab)\,(cd)]$ *is more motivated than the least motivated team in* $[(ij)\,(kl)]$ *and, in addition, the most motivated team in* $[(ab)\,(cd)]$ *is more motivated than the most motivated team in* $[(ij)\,(kl)]$*, then* $[(ab)\,(cd)]$ *is more efficient than* $[(ij)\,(kl)]$.

Figure II.2 gives an example of a social network that allows for the application of Proposition II.1. In this social network, agents $a$ and $b$ are perfectly altruistic to one another, $\delta_{ab} = 1$. They both are less altruistic toward $c$ and $d$: $\delta_{ac} = \delta_{bd} = \frac{1}{2}$ and $\delta_{ad} = \delta_{bc} = 0$. By (II.5), $a$'s and $b$'s motivation to cooperate is strongest in team $(ab)$, $m_{ab} = 1$ . Since $c$ and $d$ are also perfectly altruistic to one another, the efficient team assignment is $[(ab)\,(cd)]$.

Given the variety of feasible social networks, Proposition II.1 is rarely applicable. In many social networks, the formation of the most motivated team may leave the principal no other option but to also form an unmotivated team. When can a highly motivated team compensate for an unmotivated team?

Figure II.3 shows a social network in which no two team assignments are comparable in terms of Proposition II.1. Here, forming team $(ab)$, with perfect altruism $\delta_{ab} = 1$ and maximum motivation $m_{ab} = 1$, requires to also form the completely unmotivated team $(cd)$, with $\delta_{cd} = -1$ and $m_{cd} = 0$. Instead of $[(ab)\,(cd)]$, the principal may want to implement $[(ad)\,(bc)]$ so that coworkers have neutral preferences toward one another, $\delta_{ad} = 0 = \delta_{bc}$, and 'medium' motivation to cooperate, $m_{ad} = \frac{1}{2} = m_{bc}$. Or, the principal allows for a bit of spite, $\delta_{bd} = -\frac{1}{2}$, in order to allow for a bit of altruism, $\delta_{ac} = \frac{1}{2}$, such that $m_{bd} = \frac{1}{4}$ and $m_{ac} = \frac{3}{4}$. The question is: Can the highly motivated team $(ab)$ compensate for the non-cooperation within the unmotivated team $(cd)$? Or is separation of spiteful coworkers superior even though this implies the separation of altruistic coworkers? How does $[(ac)\,(bd)]$ compare to the alternatives?

Assume for the moment that the feasible assignments $[(ij)\,(kl)]$ do not differ in the average motivation $\frac{1}{2}\,(m_{ij} + m_{kl})$ of teams, - as in the case of Figure II.3 where this average is always 1/2. Suppose first, team output is elastic in the team members' motivation to cooperate, $\varepsilon_{x_i^*, m_{ij}} > 1$. By Lemma II.1, the principal's profit (II.7) is convex in team $(ij)$'s motivation $m_{ij}$ (and similarly for team $(kl)$). Suppose further, $(ij)$ has a higher motivation than $(kl)$, $m_{ij} \geq m_{kl}$. Increasing now $(ij)$'s motivation by $\Delta m$ overcompensates for decreasing $(kl)$'s motivation by the same quantity $\Delta m$, while keeping average motivation constant. Accordingly, the motivation asymmetry between teams should be greatest, and the principal would implement the assignment that allows for the maximum feasible motivation for one team. Due to (II.5), the principal implements the assignment which solves

$$\max_{[(ij)(kl)]} \max\{\delta_{ij}, \delta_{kl}\}\,. \tag{II.8}$$

In the example of Figure II.3, the efficient assignment is $[(ab)\,(cd)]$, since $\max\{\delta_{ab}, \delta_{cd}\} = 1 > \max\{\delta_{ac}, \delta_{bd}\} = \frac{1}{2} > \max\{\delta_{ad}, \delta_{bc}\} = 0$ .

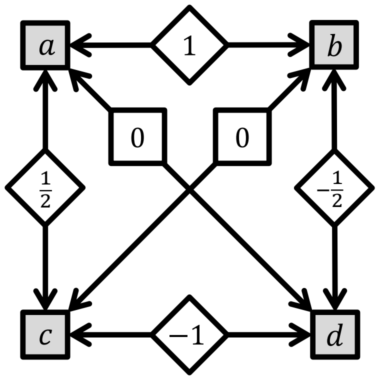


Figure II.3.: *Forming the highly motivated team* $(ab)$ *leaves the principal no other option but to also form the unmotivated team* $(cd)$.

Now suppose team output is inelastic in the team members' motivation, $\varepsilon_{x_i^*, m_{ij}} < 1$. By Lemma II.1, the principal's profit ( II.7) is concave in teams' motivations. Suppose again $m_{ij} \geq m_{kl}$. Increasing now $(ij)$'s motivation by $\Delta m$ will not compensate for decreasing $(kl)$'s motivation by $\Delta m$. In this case, the motivation asymmetry between teams should be as small as possible, and the principal would implement the assignment that allows for the maximum motivation of the least motivated team. Due to (II.5), the principal will implement the assignment which solves

$$\max_{[(ij)(kl)]} \min \{\delta_{ij}, \delta_{kl}\}. \tag{II.9}$$

In the example of Figure II.3, now assignment $[(ad)(bc)]$ is efficient, since $\min\{\delta_{ab}, \delta_{cd}\} = -1 < \min\{\delta_{ac}, \delta_{bd}\} = -\frac{1}{2} < \min\{\delta_{ad}, \delta_{bc}\} = 0$. Notice that for both elastic and inelastic team output, implementing the assignment with 'intermediate' motivations, $[(ac)(bd)]$, is always inefficient. According to ( II.8) and (II.9), the principal must go for the extremes.

So far, we have assumed that the feasible team assignments do not differ in the average motivation of the respective teams: $m_{ab} + m_{cd} = m_{ij} + m_{kl}$ for any combination of agents $\{i, j, k, l\} = \{a, b, c, d\}$. If, in this situation, $[(ab)(cd)]$ yields a higher profit than $[(ij)(kl)]$, then, by Proposition II.1, this is even more so if the motivation of one of the teams $(ab)$ and $(cd)$ is increased. Together: If $m_{ab} + m_{cd} \geq m_{ij} + m_{kl}$, and, depending on the elasticity of team output, if $[(ab)(cd)]$ solves (II.8) or (II.9), then this assignment is efficient.

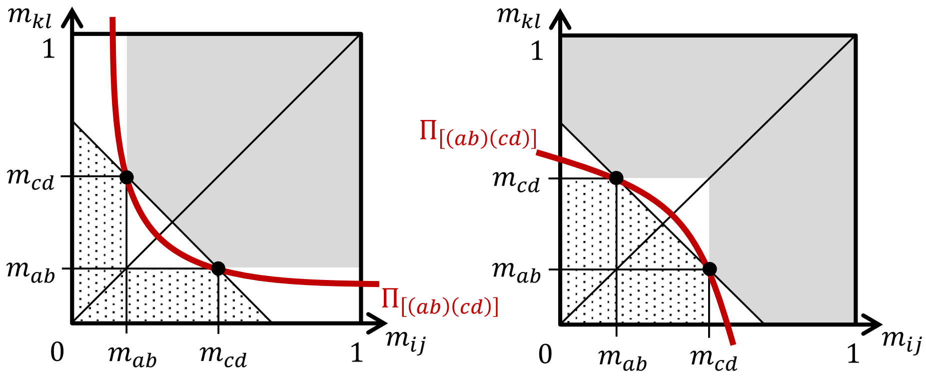


Figure II.4.: *The combinations of within-team motivation that yield more (gray) or less (dotted) profit than team assignment* $[(ab)\,(cd)]$. *With convex (concave) marginal costs of effort, the principal's isoprofit curve is convex (concave). How the white areas translate into more or less profit depends more specifically on the shape of effort cost function.*

In two $m_{ij}$-$m_{kl}$-diagrams, Figure II.4 captures the graphical analog of this reasoning. For the cases of convex (on the left) and concave (on the right) marginal cost of effort, it depicts the principal's isoprofit curve associated with the profit from a specific assignment $[(ab)\,(cd)]$. By Lemma II.1, isoprofit curves are convex (concave) if the marginal costs of effort are convex (concave). Since profit is symmetric in the teams' motivations, $(m_{ij}, m_{kl})$, isoprofit curves are symmetric to the 45°-line. Besides concavity or convexity, no further assumptions have been made about the specific shape of the function of marginal effort costs. Accordingly, the white areas in the diagrams of Figure II.4 capture exactly all those pairs $(m_{ij}, m_{kl})$ of within-team motivation that can locate above as well as below the isoprofit curve through $(m_{ab}, m_{cd})$, depending on the specific shape of marginal effort costs. Inevitably superior (inferior) to $(m_{ab}, m_{cd})$ are all those pairs of within-team motivations belonging to the gray (dotted) area. The gray and dotted areas are exactly those that cannot be crossed by any convex or concave isoprofit curve through $(m_{ab}, m_{cd})$. For convex marginal costs of effort, each combination $(m_{ij}, m_{kl})$ in the gray shaded area satisfies $m_{ij} + m_{kl} \geq m_{ab} + m_{cd}$ and $\min\{\delta_{ij}, \delta_{kl}\} \geq \min\{\delta_{ab}, \delta_{cd}\}$; and compared to every combination in the dotted area, $(m_{ab}, m_{cd})$ satisfies $m_{ab}+m_{cd} \geq m_{ij}+m_{kl}$ and $\min\{\delta_{ab}, \delta_{cd}\} \geq \min\{\delta_{ij}, \delta_{kl}\}$. On the other hand, for concave marginal costs of effort, each combination $(m_{ij}, m_{kl})$ in the gray shaded area satisfies $m_{ij}+m_{kl} \geq m_{ab}+m_{cd}$ and $\max\{\delta_{ij}, \delta_{kl}\} \geq \max\{\delta_{ab}, \delta_{cd}\}$; and compared to every combination in the dotted area, $(m_{ab}, m_{cd})$ satisfies $m_{ab}+m_{cd} \geq m_{ij}+m_{kl}$ and $\max\{\delta_{ab}, \delta_{cd}\} \geq \max\{\delta_{ij}, \delta_{kl}\}$. This proves the following Proposition.

**Proposition II.2** *Suppose average motivation in* $[(ab)(cd)]$ *is at least as high as in* $[(ij)(kl)]$*. Then* $[(ab)(cd)]$ *is more efficient than* $[(ij)(kl)]$ *if one of the following conditions is satisfied.*

*(i) Marginal costs of effort are concave, and the most motivated team in* $[(ab)(cd)]$ *is more motivated than the most motivated team in* $[(ij)(kl)]$*.*

*(ii) Marginal costs of effort are convex, and the least motivated team in* $[(ab)(cd)]$ *is more motivated than the least motivated team in* $[(ij)(kl)]$*.*

Proposition II.2 suggests a rule of thumb that relates efficient team assignment to the steepness of the effort cost function ($C_{xxx} \gtrless 0$ ): 'The more difficult the production of any additional unit of team output is the more important is the separation of agents who are less altruistic to one another.'

## II.4. Pooling and the Integration of a Spiteful Outcast

To assign teams efficiently, the principal needs the information on the agents' social network. Obtaining this information might be costly. Can the principal 'get around' the team assignment decision by simply paying each agent an equal share of the market return on overall output? Even though each agent would still have to collaborate with some other agent in a team of two, he would internalize the externalities of his efforts on all the other agents, regardless of whom he is sitting in the same boat with. In terms of incentives, payoffs, and profits, actual team assignment could then be arbitrary. Intuitively, one might expect that this cannot be more efficient than each of the feasible team assignments $[(ij)(kl)]$, where teams are paid for their own team output: Receiving an equal share of the overall return would increase every agent's incentive to free-ride; the group of beneficiaries of individual effort should thus be as small as practically feasible. The objective of the following analysis is to confirm this intuition for most social networks, but to reject it for some.

For some fixed piece rate $w$, let $[(abcd)]$ denote the principal's strategy to pay each agent an equal share, $\frac{1}{4}w$, of overall return. I refer to $[(abcd)]$ as *the pooling of incentives* or, likewise, as 4-*assignment*. By contrast, I refer to the assignments $[(ab)(cd)]$, $[(ac)(bd)]$, and $[(ad)(bc)]$ as $2+2$ -*assignments*.

Consider some combination of agents $\{i,j,k,l\} = \{a,b,c,d\}$, and suppose the principal utilizes the pooling of incentives. Then, with all else equal, agent $i$ ends up with material wealth

$$\pi_i = \frac{1}{4} w\left(x_a + x_b + x_c + x_d\right) - C\left(x_i\right). \tag{II.10}$$

Since within-team efforts are perfectly substitutable, this holds regardless of who teams up with whom exactly. Again, $i$ maximizes utility $u_i = \pi_i + \delta_{ij}\pi_j + \delta_{ik}\pi_k + \delta_{il}\pi_l$ with respect to individual effort $x_i$. His marginal utility from exerting effort is given by

$$\frac{du_i}{dx_i} = \left(\frac{1}{4} + \frac{1}{4}\delta_{ij} + \frac{1}{4}\delta_{ik} + \frac{1}{4}\delta_{il}\right) w - C_x\left(x_i\right). \tag{II.11}$$

Marginal return on effort, $\left(\frac{1}{4} + \frac{1}{4}\delta_{ij} + \frac{1}{4}\delta_{ik} + \frac{1}{4}\delta_{il}\right) w$, reflects the fact that now only a quarter of $i$'s individual effort transforms into effective return for $i$. The remaining three quarters are in favor of the other agents. The weight $\frac{1}{4} + \frac{1}{4}\delta_{ij} + \frac{1}{4}\delta_{ik} + \frac{1}{4}\delta_{il}$ measures $i$'s valuation of this income effect. I refer to

$$M_i = \frac{1}{4} + \frac{1}{4}\delta_{ij} + \frac{1}{4}\delta_{ik} + \frac{1}{4}\delta_{il} \tag{II.12}$$

as *$i$'s motivation in* 4-*assignment* $[(abcd)]$. This motivation increases in $i$'s altruism toward each of the other agents. Obviously, $M_i \in \left[-\frac{1}{2}, 1\right]$. If $i$'s motivation in $[(abcd)]$ is negative, he will exert zero effort, since every unit of effort then yields negative marginal return. Therefore, $i$ has a dominant strategy in exerting effort

$$x_i^* = C_x^{-1}\left(w \max\left\{0, M_i\right\}\right). \tag{II.13}$$

The principal's profit associated with the pooling of incentives is given by

$$\Pi_{[(abcd)]} = \sum_{i\in\{a,b,c,d\}} (1-w)\, C_x^{-1}\left(w \max\{0, M_i\}\right). \tag{II.14}$$

Not surprisingly, 4-assignment $[(abcd)]$ does not in general yield a higher profit than every $2+2$-assignment. An obvious example is the social network of Figure II.2. There, the efficient $2+2$-assignment is $[(ab)(cd)]$ in which every agent exerts an effort $C_x^{-1}(w)$ . Due to the symmetry of this social network, the agents are identically motivated under the pooling of incentives; namely, $M_i = \frac{1}{4}\left(1+1+\frac{1}{2}+0\right) = \frac{3}{8}$ for all $i$. By (II.13), agents each exert an effort $C_x^{-1}\left(\frac{3}{8}w\right) < C_x^{-1}(w)$. Thus, $\Pi_{[(abcd)]} < \Pi_{[(ab)(cd)]}$. In this example, the inferiority of $[(abcd)]$ is driven by the decrease of every agent's motivation when enlarging the group of beneficiaries of their individual efforts.

Suppose $i$'s motivation in $[(abcd)]$ is positive, $M_i > 0$. As before, $i$'s optimum individual effort is elastic in $i$'s motivation to cooperate, $\varepsilon_{x_i^*,M_i} > 1$, if the marginal costs of effort are concave. Optimum effort is inelastic, $\varepsilon_{x_i^*,M_i} < 1$, if the marginal costs of effort are convex. The intuition behind Proposition II.2 raises the question: When implementing $[(abcd)]$ instead of the efficient $2+2$-assignment, would the potential increase in some agents' motivation compensate for another agent's demotivation? The answer is: Yes, sometimes. For this, the marginal costs of effort must be convex, and the respective social networks must contain an agent who can be named a 'spiteful outcast'.

I refer to agent $d$ as a *spiteful outcast* in the social network $\{a,b,c,d\}$ if $d$ is completely unmotivated in $[(abcd)]$, $M_d < 0$, while all the other agents' motivation in $[(abcd)]$ is positive: $M_a, M_b, M_c > 0$. The term 'spiteful outcast' is suggested by the following observation: The definition implies that $\delta_{ad}+\delta_{bd}+\delta_{cd} < -1$ and $\delta_{ad}+\delta_{bd}+\delta_{cd} < \delta_{ab}+\delta_{ac}+\delta_{bc}$.[12] Hence, agents $\{a,b,c\}$ are on average spiteful toward $d$, a spite that is reciprocated by $d$, and their average spite toward $d$ is stronger than their average spite toward one another (if they are spiteful toward one another at all). Relative to their relationships to $d$, agents $\{a,b,c\}$ form a *clique* that dislikes and is disliked by the spiteful outcast $d$. The following

[12]Let $M_d < 0$, and $-M_a, -M_b, -M_c < 0$. Adding all four inequalities yields $-\frac{1}{2}\left(1+\delta_{ab}+\delta_{ac}+\delta_{bc}\right) < 0$. Thus, $\delta_{ad}+\delta_{bd}+\delta_{cd} < -1 < \delta_{ab}+\delta_{ac}+\delta_{bc}$.

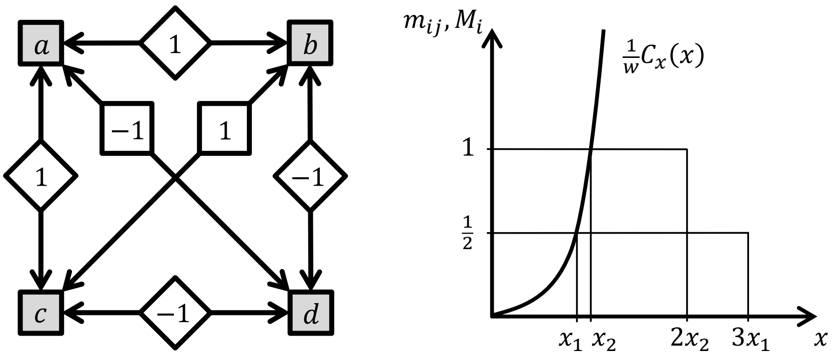


Figure II.5.: *A social network with a spiteful outcast, d, and a production process for which pooling the incentives is efficient. In $[(abcd)]$, agents a, b, and c each exert an effort $x_1$, while d exerts zero effort. In $[(ab)(cd)]$, a and b each exert an effort $x_2$, while c and d exert zero effort.*

example shows that integrating a spiteful outcast through the pooling of incentives can indeed be more efficient than any $2+2$-assignment.

Figure II.5 depicts a clique of agents $\{a,b,c\}$ who are perfectly altruistic toward one another. Each member is perfectly spiteful toward agent $d$, and vice versa. The vector $(M_a, M_b, M_c, M_d)$ of the agents' motivations in 4-assignment $[(abcd)]$ equals $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\right)$. By definition, $d$ is a spiteful outcast. All the $2+2$-assignments are symmetric: in each, there is one highly motivated team and one completely unmotivated team. Cooperation between $d$ and a clique member $i \in \{a,b,c\}$ is zero in any $2+2$ -assignment: $m_{id} = 0$. Now focus on $[(ab)(cd)]$. By (II.6), clique members $a$ and $b$ each exert an effort $C_x^{-1}(w)$, while agents $c$ and $d$ each exert zero effort. The pooling of incentives increases $c$'s motivation (from $m_{cd} = 0$ to $M_c = \frac{1}{2}$) at the expense of demotivating agents $a$ and $b$ (from $m_{ab} = 1$ to $M_a = M_b = \frac{1}{2}$). Can it be that the increase in $c$'s motivation overcompensates for the decrease in $a$'s and $b$'s motivation? - The answer depends on the shape of the effort cost function.

Figure II.5 also depicts a production process, represented by the marginal costs of effort, that indeed makes the pooling of incentives the best response to the underlying social network. Here, team output is inelastic in the motivation to cooperate, $C_{xxx} > 0$. In $[(abcd)]$, each clique member exerts an effort $x_1 = C_x^{-1}\left(\frac{1}{2}w\right)$. Overall output is $3x_1$, since $d$ exerts zero effort. In $[(ab)(cd)]$, clique members $a$ and $b$ each exert an effort $x_2 = C_x^{-1}(w)$

while $c$ and $d$ exert zero effort. The marginal costs of effort are chosen such that, when implementing $[(abcd)]$ instead of $[(ab)\,(cd)]$, the motivation increase of $c$ overcompensates for the demotivation of $a$ and $b$: $3x_1 > 2x_2$. This positive effect of integrating a spiteful outcast is not driven by the outcast himself, who is demotivated anyway; motivating clique member $c$ renders integration of $d$ efficient.[13]

Notice that the condition $C_{xxx} > 0$ does not suffice to make the integration of a spiteful outcast efficient: If $C_x$ in Figure II.5 was chosen nearly linear, the positive effect of integration would collapse.

Of course, the social network of Figure II.5 would allow for an even more efficient payment scheme: pay each member of the clique $\{a, b, c\}$ an equal share of overall return, and pay $d$ zero. But doing so requires the principal to be informed about the agents' network. The consideration of pooling rather serves the purpose of understanding whether the principal can avoid obtaining such information. Even though this is only occasionally the case, I discuss the pooling of incentives in detail because the findings reject the naive intuition that pooling would never be more efficient than incentivizing the agents on a smaller group level.

**Proposition II.3** *The pooling of incentives can be more efficient than any* 2+2*-assignment if output is inelastic in the agents' motivation to cooperate and, at the same time, the social network contains a spiteful outcast. Otherwise, there exists at least one* $2+2$*-assignment that is more efficient than pooling.*

**Proof.** See Appendix B. ■

In the absence of a spiteful outcast, or if the marginal costs of effort are concave, teams (in terms of compensation) should be as small as practically feasible. The imperative of

[13]The definition of a spiteful outcast characterizes the efficiency of pooling in the following sense:

(1) Without a spiteful outcast, pooling is always inferior (see Proposition II.3).

(2) Pooling can be efficient even if $M_a, M_b, M_c > 0$ are arbitrarily small: Let $\delta_{ab}, \delta_{ac}, \delta_{bc} = 5\varepsilon$, and $\delta_{ad}, \delta_{cd}, \delta_{bd} = 2\varepsilon - 1$. Then, $M_d = 3\varepsilon - \frac{1}{2}$; $M_a, M_b, M_c = 3\varepsilon$; $m_{ab} = \frac{1}{2} + 5\varepsilon$; $m_{cd} = \varepsilon$. For any $\varepsilon > 0$ sufficiently small, marginal effort costs can be chosen appropriately (similarly as in Figure II.5), such that pooling is efficient.

(3) Pooling can be efficient even if $M_d < 0$ is arbitrarily close to zero: Let $\delta_{ab}, \delta_{ac}, \delta_{bc} = 1$, and $\delta_{ad}, \delta_{cd}, \delta_{bd} = -\frac{1}{3} - \varepsilon$. Then, $M_d = -\frac{3}{4}\varepsilon$; $M_a, M_b, M_c = \frac{2}{3} - \frac{3}{4}\varepsilon$; $m_{ab} = 1$; $m_{cd} = \frac{1}{3} - \frac{1}{2}\varepsilon$. For any $\varepsilon > 0$ sufficiently small, marginal effort costs can be chosen appropriately, such that pooling is efficient.

small teams imposes an information problem on the principal. Even for assessing whether or not pooling can be efficient, she needs to have the information on the interpersonal relationships at work.

## II.5. Delegation or Control?

Suppose the principal is uninformed about the social network of her staff while, as before, preferences are common knowledge among agents. Can the principal leave the decision on who teams up with whom to her staff?

One can think of several ways of how this decision could be 'delegated'. One is to 'let the agents vote', according to some well-designed voting rules. Another one is to make the agents reveal their interpersonal degrees of altruism by applying some more sophisticated mechanism. I focus here on the first of these alternatives. I do not address the question of how the principal would incentivize the agents to internalize the externalities that their actions impose on her, which I interpret as 'control'.[14]

I consider *delegation mechanisms*, by which I refer to strategic games between the agents that translate the announced preferences for (or preference orderings over) team compositions into final team assignments but that do not affect the agents' ex post utilities beyond team assignments. Without loss of generality, side payments to or between agents can be neglected.

Denote by $\mathbb{A} = \{[(ab)(cd)], [(ac)(bd)], [(ad)(bc)], [(abcd)]\}$ the set of feasible team assignments. For $i \in \{a, b, c, d\}$, denote by $T$ the set of (every) agent $i$'s feasible preference orderings over $\mathbb{A}$. Suppose the social network determinants $(\delta_{ab}, \delta_{ac}, \delta_{ad}, \delta_{bc}, \delta_{bd}, \delta_{cd}) \in [-1, 1]^6$ are common knowledge among the agents, but unknown to the principal.[15] The agents thus know which team assignment $A^* \in \mathbb{A}$ maximizes the principal's profit, the principal herself does not.[16]

[14]The adequate tool for the latter approach is *Nash implementation* in the manner of Maskin (1999) and Moore and Repullo (1990).

[15]To be sure, the principal has a belief about the social network structure realized in her staff, represented by a probability density $f : [-1, 1]^6 \to \mathbb{R}$ on the set of feasible combinations of degrees of altruism. In what follows, the essential assumption is that $f(x) > 0$ for all $x \in (-1, 1)^6$.

[16]For the purpose of this Section, it suffices to concentrate on social networks for which the profit maximizer $A^*$ is unique.

The principal seeks to *ensure* implementation of $A^*$ by application of a mechanism as follows. She asks agents simultaneously to tell her their true preference orderings over $\mathbb{A}$ and builds her decision to implement a team assignment $A \in \mathbb{A}$ upon the collection $[t_a, t_b, t_c, t_d] = t \in T^4$ of messages received.

A *delegation mechanism* M is a collection of density functions $\mathrm{m}[t] : \mathbb{A} \to [0,1]$ which, conditional on the messages received, determine the probabilities with which every $A \in \mathbb{A}$ will be implemented. Thus, $\sum_{A \in \mathbb{A}} \mathrm{m}[t](A) = 1$ for each $t \in T^4$. The principal seeks to design M in such a way that equilibrium behavior under M yields a profit maximizing team assignment with likelihood 1.

Denote by $\tau_i : T \to [0,1]$ an agent $i$'s (possibly degenerate) mixed strategy which, for every $t_i \in T$, determines the probability $\tau_i(t_i)$ with which $i$ announces that his preference ordering over $\mathbb{A}$ was $t_i$. Let $u_i(A)$ denote $i$'s utility from implementation of $A \in \mathbb{A}$. Agent $i$'s expected utility associated with the mixed strategy profile $(\tau_i, \tau_{-i})$ of all agents is thus given by

$$EU_i(\tau_i, \tau_{-i}) = \sum_{[t_i, t_{-i}] \in T^4} \prod_{j \in \{a,b,c,d\}} \tau_j(t_j) \left( \sum_{A \in \mathbb{A}} \mathrm{m}[t_i, t_{-i}](A) \cdot u_i(A) \right), \tag{II.15}$$

where $\tau_{-i}$ and $t_{-i}$, respectively, collect the mixed strategies of and the messages from the agents other than $i$. A mixed strategy Nash equilibrium of M is a collection $[\tau_a^*, \tau_b^*, \tau_c^*, \tau_d^*]$ of mixed strategies such that $EU_i(\tau_i^*, \tau_{-i}^*) \geq EU_i(\tau_i, \tau_{-i}^*)$ for all $\tau_i$ available to each $i \in \{a,b,c,d\}$.

As an example, consider the way teachers tend to form soccer teams during sports classes: The teacher announces two students who then, by turns, select their preferred class mates out of those still unselected. Within the model framework, the following mechanism $\overline{\mathrm{M}}$ provides agents $\{a,b,c,d\}$ with equivalent incentives. The principal randomly assigns one agent $i$ who then selects his preferred team mate and, thereby, dictates the composition of the remaining team; agent $i$ then decides whether every team is to be paid for their own team output or whether the incentives are to be pooled. That is, $\overline{\mathrm{m}}[t_i', t_{-i}](A') = 1$ for all $t_{-i}$ if and only if assignment $A'$ is most preferred by $i$ according to the announced preference ordering $t_i'$. In this case, agent $i$ has a weakly dominant strategy in announcing one of those preference orderings in $T$ that rank his truly preferred assignment first. The mixed strategies $\tau^*_{-i}$ of the agents other than $i$ can be chosen arbitrarily, since they do not affect the outcome.

In the following, I show that team assignment that is unanimously preferred by all the agents *might not* maximize the principal's profit. As soon as the social network is asymmetric, subjective well-being on the one hand, and material efficiency on the other, can collide. This potential divergence negates the existence of a delegation mechanism that *ensures* maximum overall productivity and maximum profit for the principal.

The results of Sections II.3 and II.4 indicate that the uninformed principal cannot rule out any of the four team assignments in $\mathbb{A}$ when she seeks to maximize her profit. Since she wants to *ensure* maximum profit, M must not preclude any feasible team assignment from being implemented with certainty. Formally:

**Condition II.1** *For any $A' \in \mathbb{A}$, there exists at least one collection of messages $t' \in T^4$ for which the respective density function $\mathrm{m}[t']$ satisfies $\mathrm{m}[t'](A') = 1$.*

Condition II.1 implies that, if there is a social network for which a specific team assignment $A' \in \mathbb{A}$ is Pareto dominant for the group of agents, then M implements $A'$ in (some) Nash equilibrium. But with $A'$ being Pareto dominant, and given (II.15), the following strategy profile does constitute a Nash equilibrium: If $t'$ satisfies $\mathrm{m}[t'](A') = 1$, then $\tau_i(t'_i) = 1$ for each agent $i$.

**Proposition II.4** *There is no delegation mechanism that, for any social network, ensures the profit maximizing composition of teams.*

**Proof.** In light of Condition II.1, it suffices to identify a social network that provides all the agents with a unanimous preference for an assignment $A' \neq A^*$.

Consider the social network in Figure II.6. For $\beta, \gamma, \delta \in (-1, 1)$, assume $\beta > \gamma > \delta$. According to Proposition II.1, $A^* = [(ab)(cd)]$. Behold the symmetry of this network. Let $\pi_\beta$, $\pi_\gamma$, and $\pi_\delta$ denote every agent's material wealth in $[(ab)(cd)]$, $[(ac)(bd)]$, and $[(ad)(bc)]$ , respectively. Let $\pi_4$ denote every agent's material wealth in 4 -assignment $[(abcd)]$. Every agent $i$ then realizes utility $u_i = (1 + \beta + \gamma + \delta)\pi_\beta$ in $[(ab)(cd)]$, utility

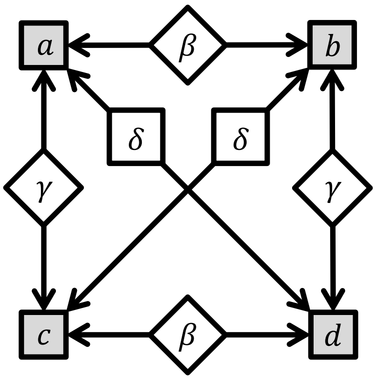


Figure II.6.: *If $\beta > \gamma > \delta$, the efficient team assignment is $[(ab)\,(cd)]$. If $\beta+\gamma+\delta < -1$, however, the agents unanimously prefer $[(ad)\,(bc)]$. Each delegation mechanism would implement $[(ad)\,(bc)]$.*

$u_i = (1+\beta+\gamma+\delta)\,\pi_\gamma$ in $[(ac)\,(bd)]$, utility $u_i = (1+\beta+\gamma+\delta)\,\pi_\delta$ in $[(ad)\,(bc)]$, and utility $u_i = (1+\beta+\gamma+\delta)\,\pi_4$ in $[(abcd)]$. Notice that $\pi_\beta > \pi_\gamma > \pi_\delta > 0$. Now assume $\beta+\gamma+\delta < -1$. Thus, $M_i < 0$ for each $i$ and, by (II.13), $\pi_4 = 0$. In this case, each agent $i$'s preference ordering is given by

$$[(abcd)] \succ_i [(ad)\,(bc)] \succ_i [(ac)\,(bd)] \succ_i [(ab)\,(cd)]\,. \tag{II.16}$$

Hence, $A^*$ is least preferred: The desired mechanism does not exist. ■

The counter examples presented in this proof are not just artifacts; they do not involve the effort cost function and its effects on equilibrium effort choices. Furthermore, since all inequalities in the proof of Proposition II.4 are strict, it is easy to see that the counter examples hold when adding a little, independently distributed noise to each degree of altruism in the network of Figure II.6. Nevertheless, these social networks might appear unrealistic, or ‘unlikely’: Since $\beta+\gamma+\delta < -1$, they all contain at least two dyads engaged in conflict (spite).

However, it is not necessarily spite that rejects the existence of a delegation mechanism yielding efficient team assignment: Loosely speaking, it is rather the social network's (potential) asymmetry that rejects the existence of such a mechanism. In the absence of spite, this asymmetry concerns configurations of the form $\delta_{ab} > \delta_{ac}, \delta_{ad}, \delta_{bc}, \delta_{bd} > \delta_{cd}$.

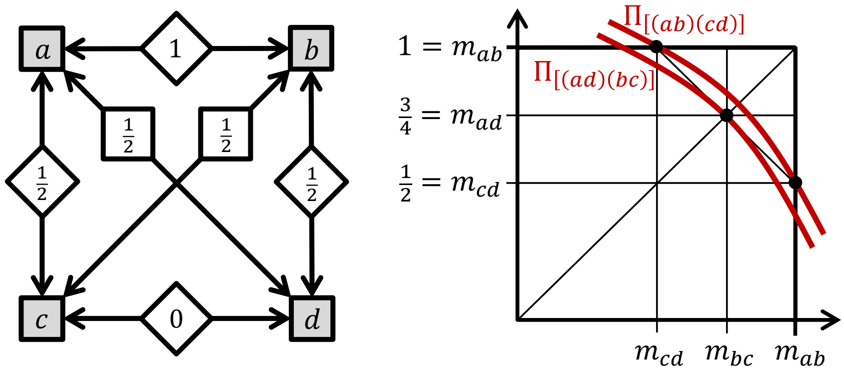


Figure II.7.: *A social network in which no two agents are spiteful toward one another and a production process for which marginal effort costs and thus the principal's isoprofit curves are concave but nearly linear. The agents unanimously prefer* $[(ac)(bd)]$*, or* $[(ad)(bc)]$*, while the principal prefers* $[(ab)(cd)]$*.*

**Proposition II.5** *Even in the absence of spite, the social network and the costs of effort can be such that agents unanimously prefer a team assignment that yields inefficient production overall.*

**Proof.** Consider the social network of Figure II.7. By Proposition II.3, $[(abcd)]$ need not be considered to justify the argument. Notice that $[(ac)(bd)]$ and $[(ad)(bc)]$ are equivalent by symmetry. Suppose $C_{xxx} < 0$. By Proposition II.2, the principal prefers $[(ab)(cd)]$.

Choose $C_x(x) = x^{1/\kappa}$, with $\kappa > 1$. Thus, $C_x^{-1}(m_{ij}w) = m_{ij}^{\kappa}w^{\kappa}$. Then each member of a team $(ij)$ realizes material wealth

$$\pi_i(m_{ij}) = \frac{w^{1+\kappa}}{1+\kappa} m_{ij}^{\kappa}(1+\kappa-\kappa m_{ij}), \tag{II.17}$$

which increases in $m_{ij} \in [0,1)$. Motivations would be $m_{ab} = 1$ and $m_{cd} = \frac{1}{2}$ in $[(ab)(cd)]$, and $m_{ij} = \frac{3}{4}$ in all the other $2+2$-assignments. Comparison of the respective utilities from team assignment reveals that each agent prefers $[(ac)(bd)]$ as $\kappa \to 1$: In $[(ab)(cd)]$, agents $a$ and $b$ each realize utility

$$u_{[(ab)(cd)]}^{a,b} = \frac{w^{1+\kappa}}{1+\kappa}\left[2+\left(\frac{1}{2}\right)^{\kappa}\left(1+\frac{\kappa}{2}\right)\right] \xrightarrow{\kappa\to 1} \frac{w^{1+\kappa}}{1+\kappa}\cdot\frac{11}{4}, \tag{II.18}$$

and in $[(ac)(bd)]$, they each realize utility

$$u^{a,b}_{[(ac)(bd)]} = \frac{w^{1+\kappa}}{1+\kappa} 3 \left(\frac{3}{4}\right)^{\kappa} \left(1 + \frac{\kappa}{4}\right) \xrightarrow{\kappa \to 1} \frac{w^{1+\kappa}}{1+\kappa} \cdot \frac{45}{16} > \frac{w^{1+\kappa}}{1+\kappa} \cdot \frac{11}{4}. \tag{II.19}$$

In $[(ab)(cd)]$, agents $c$ and $d$ each realize utility

$$u^{c,d}_{[(ab)(cd)]} = \frac{w^{1+\kappa}}{1+\kappa} \left[1 + \left(\frac{1}{2}\right)^{\kappa} \left(1 + \frac{\kappa}{2}\right)\right] \xrightarrow{\kappa \to 1} \frac{w^{1+\kappa}}{1+\kappa} \cdot \frac{7}{4}, \tag{II.20}$$

and in $[(ac)(bd)]$, they each realize utility

$$u^{c,d}_{[(ac)(bd)]} = \frac{w^{1+\kappa}}{1+\kappa} 2 \left(\frac{3}{4}\right)^{\kappa} \left(1 + \frac{\kappa}{4}\right) \xrightarrow{\kappa \to 1} \frac{w^{1+\kappa}}{1+\kappa} \cdot \frac{30}{16} > \frac{w^{1+\kappa}}{1+\kappa} \cdot \frac{7}{4}. \tag{II.21}$$

Finally, since motivations (II.5) in $[(ac)(bd)]$ are greater than motivations (II.12) in $[(abcd)]$, and since all agents attach positive weight to the wealth of (almost) all other agents, $[(ac)(bd)]$ Pareto dominates $[(abcd)]$.

Notice that even though marginal effort costs are concave, individual material wealth from teamwork (II.17) is not convex but concave in $m_{ij}$ if $\kappa > 1$ is sufficiently small: Since $m_{ij} \in \left[\frac{1}{2}, 1\right]$, we have $\frac{d^2}{dm_{ij}^2}\pi_i = w^{1+\kappa} k m_{ij}^{k-2} \left[k\left(1 - m_{ij}\right) - 1\right] < 0$ for all $\kappa \in (0, 2)$. Material wealth from team output is therefore inelastic in a team's motivation. With nearly linear marginal costs of effort ($\kappa \to 1$), the gains of $c$ and $d$ when implementing $[(ac)(bd)]$ instead of $[(ab)(cd)]$ outweigh the respective losses of $a$ and $b$. Social preferences are such that $a$ and $b$ are willing to sacrifice some of the returns from their joint teamwork in order to support $c$ and $d$, and $c$ and $d$ prefer $a$ and $b$ to make that sacrifice. ■

Again, the example is robust with respect to the addition of noise to the agents' degrees of altruism, and by the same principle, many more such examples can be constructed.

It would thus be naive to think that, just because all group members value each other's material well-being, team assignment that is unanimously preferred by all group members would coincide with efficient overall production.

## II.6. Conclusion

Efficient team assignment is sensitive to the determinants of the agents' social network of interpersonal relationships as well as to the shape of individual effort costs. If the marginal costs of effort are convex, team output responds inelastically to an increase in team members' willingness to cooperate; in this case, the principal must focus on separating the least cooperative groups of agents. Conversely, if the marginal costs of effort are concave, team output responds elastically; and the principal must focus on grouping those agents which cooperate most. In any case, team assignment in the agents' social network poses a problem of information to the principal. Neither the pooling of incentives, nor the delegation of team assignment ensure overall efficient production. Even in the absence of spite, agents might unanimously opt for a composition of teams that makes production inefficient. The principal has no option but to acquire information on the interpersonal relationships in the workplace, and intervene in the team assignment process when necessary. The gains or savings acquired from staffing under the awareness of the agents' social network may outweigh the hidden costs of control (Falk and Kosfeld, 2006) and, in the case of intervention, the forgone motivational effects of delegation (Charness et al., 2012).

The old-fashioned way of observing the interpersonal relationships at work is to spend time with personnel: watching them, talking to them, observing who joins whom for lunch. But having entered the era of 'big data', another option has evolved. A market has emerged in which firms offer the investigation of individuals' behavior patterns within organizations. These analyses involve linguistics and utilize data traces that employees leave whenever communicating digitally.[17] These strategies of information acquisition are costly. No matter the approach, they require time or cause expenses. This study indicates that staffing managers might have a willingness to pay for information on their employees' interpersonal relationships.

[17]For a brief overview see Hoffmann (2010); for a detailed, popular description see Charnock (2010).

# Chapter III.

# Parenting and Law Enforcement: On the Determination of People's Ethics of Law-Abidance[1]

## III.1. Introduction

The extent to which people abide by laws is generally believed to be influenced, if not determined, by the interplay of the following factors: the returns from non-compliance, the threats of formal law enforcement, social norms, and people's ethical convictions regarding law abidance.[2] A sound understanding of this interplay is necessary to render formal law enforcement effective, or even efficient.

While most of the related literature has focused on the determinants and evolution of social norms of law abidance, we, in this study, are concerned with the determinants of people's ethics of law abidance. We focus on an institution that we consider natural when it comes to shaping those ethics; namely, the family. We ask how a family's expected intergenerational economic standing might incentivize parents to bring up their children as law-abiding citizens, and how these incentives are affected by formal law enforcement.

[1]This chapter is based on an unpublished working paper co-authored by Aart Gerritsen and Vai-Lam Mui. Naturally, I am responsible for any errors.

[2]Their ethics of law abidance provide people with an *intrinsic* motivation to abide by the law, whereas formal law enforcement and social norms provide them with *extrinsic* motivations.

Our analysis is based on the following rationale. Grown up children rationally decide on the extent to which they engage in illegal activities (e.g., tax evasion, corruption, theft, or doing business in the shadow economy). Altruistic parents are likely to financially support their children when those are in need. They anticipate that, if their non-compliant children are caught and convicted, they might well end up bearing part of the (financial) burden of punishment. A rational child, in turn, can count on parental support in case of conviction and therefore engages even more in illegal activities. Their partial insurance provision thus imposes a moral hazard problem on parents. This variant of the 'Samaritan's dilemma' (Buchanan, 1975, and Bruce and Waldman, 1990) provides parents with the incentive to influence their adult children's behavior by instilling in them an ethic of law abidance in advance, while they are adolescent. However, altruistic parents also care about the material well-being of their children, which is harmed by the ethics of law-abidance through their distortion of children's incentives. Parents thus face a trade-off between improving their own well-being and that of their children.

Whether parents are exposed to this variant of the 'Samaritan's dilemma' depends on their expectations about their children's legal (and illegal) income prospects—as compared to their own wealth: If expectations are low, parents anticipate they would have to support their children in any case, and thus benefit ex ante from their children's illegal efforts through a reduction in financial support; if expectations are high, financial support would never be provided, regardless of conviction. The described incentive of parents to bring up their children as law-abiding citizens therefore depends on expected intergenerational social mobility.

At this point, some of our readers might wonder what kind of families, or laws, we are talking about, as they may not recognize the above considerations as relevant for themselves—neither as parents nor as their parents' children. Two examples help clarify the scope of our analysis.

*"Grub first, then ethics"* (Berthold Brecht): With regard to the socio-economic environment of a family, the above described parent-child relationship might not be representative for families who are located in the suburban middle-class neighborhoods of a developed

country, which are arguably the ones that law enforcement authorities are *least* concerned with. Parents in such environments might try to foster their children's law abidance simply because they consider crime to be "bad". Instead, law enforcement authorities are much more concerned, and so are we, with those socio-economic environments in which illegitimate activities are an essential part of a community's everyday business and *might* contribute substantially to a person's regular income. We thus think of parent-child relationships that are located in impoverished, under-class communities.[3]

The criminological literature on 'gang criminality' supports our view that, in those communities, some parents do take into account the material incentives related to crime when bringing up their children, and that these incentives might eventually influence adult children's criminality. For instance, Anderson (1999, p.133) interviewed families in impoverished US inner cities. He reports that some parents, though inwardly disapproving their sons' involvement in juvenile delinquency, saw gang membership as an 'opportunity' for their sons to become financially independent, and tacitly accepted it by 'turning a blind eye'.[4] Psychological studies, on the other hand, have shown that 'behavioral control', loosely speaking the opposite of 'turning a blind eye', is a key factor for parents to transmit their own values to their children (e.g., Barber et al., 2005). In particular, McCord (1991), in a longitudinal study with a cohort of US citizens, has shown that a lack of 'behavioral control' by parents is positively correlated with juvenile delinquency, while juvenile delinquency is a good predictor of adult criminality.[5]

*Government's law as an antagonist to people's ethics*:[6] Whether parents consider "crime" to be "bad" depends on whether they agree with the law in the first place. Parents might perceive government's law as unjust, unsubstantiated, or they might simply despise the regime that intends to govern them (as it has often been the case in countries that were conquered by another nation). Parents who do not agree with the norms enforced by law

[3]See Wilson (1997) and MacDonald and Marsh (2005) for a discussion of the challenges faced by youths and parents in such communities.

[4]For a literature survey on the role of the family in young people's gang involvement see Young, Fitzgibbon, and Silverstone (2014).

[5]Similar results have been obtained for other countries; for instance, by Farrington (2003) for the UK.

[6]We adopt, and adapt, from Posner (1997) the notion of social norms as an antagonist to law.

might raise their children to disagree with those norms as well, and thus value their children's risk taking when disobeying the respective laws (and insure them partially against being convicted and punished). In such cases, parents' motivations to still bring up their children as law-abiding citizens are at most instrumental, rather than intrinsic.

Our analysis reveals a non-monotonic relationship between expected intergenerational social mobility and parents' incentives to bring up their children as law-abiding citizens. Under intergenerational *downward* mobility, incentives are *weak*, as parental support would have to be provided regardless of whether children succeeded in their illegal activities or were convicted and suffered from hefty fines. Ex ante, parents then benefit materially from their children's noncompliance through a reduction in financial support. By contrast, under intergenerational *stagnation*, incentives are *strong*, as financial support would be provided only in case of conviction, imposing the 'Samaritan's dilemma' on parents. Under intergenerational *upward* mobility, incentives are *moderate*, as parents would not support their children even if those were convicted; in this case, their mere incentive to instill an ethic of law abidance in their children is to limit the shame (if any) their children might bring upon them.

We then ask how changes in formal law enforcement would affect ethics formation. We find that, with strong intergenerational upward or downward mobility, the ethics of law abidance are invariant to more surveillance or tougher punishment. Under intergenerational stagnation as well as weak upward or downward mobility, more surveillance substitutes for parents' need to fight off their 'Samaritan's dilemma' and thus crowds out ethics formation; however, we find that the effect of tougher punishment on ethics formation is ambiguous.

The chapter proceeds as follows. Section III.2 discusses the related literature. Section III.3 outlines the basics of our model and derives the equilibrium ethics of law abidance. Section III.4 interprets results with regard to intergenerational social mobility. Section III.5 derives the comparative statics of ethics formation with respect to formal law enforcement. Section III.6 concludes.

## III.2. Related Literature

Our study relates to several strands of literature. First of all, we contribute to the literature on family economics, pioneered by Becker (1974, 1976). This literature typically investigates how the conflicting preferences within a family can lead to efficient or inefficient decision making and within-family allocations of resources. Indeed, the parents in our economic model face a particular variant of the 'Samaritan's dilemma' (Buchanan, 1975, and Bruce and Waldman, 1990); children might take advantage of their parents' partial insurance provision against being convicted by engaging too much in illegal activities. A recent strand in family economics is concerned with the incentives and means of parents to control their children's actions, either by providing them with extrinsic incentives (e.g., Weinberg, 2001), or by manipulating their preferences (e.g., Lindbeck and Nyberg, 2006, and Becker, Murphy, and Spenkuch, 2016; in the study of Doepke and Zilibotti, 2017, parents do both). We add to this literature a novel dimension, namely parents' incentives to shape their offspring's ethical convictions regarding law abidance.[7]

A growing literature is dedicated to the determinants and evolution of people's social norms and ethical convictions. Pioneered by Bisin and Verdier (2001), this strand studies how norms or ethics are transmitted vertically, from one generation to another, and (in some studies) even horizontally, meaning that there is some extent of strategic interaction between the "old" when transmitting their values to the "young". Part of this literature focuses specifically on the role of parents in the transmission of social norms and ethics, while other studies are dedicated to the interplay between cultural transmission

[7]Our formal setup is quite similar to the one of Lindbeck and Nyberg (2006), briefly L&N, who analyze parents' incentives to instill work ethics in their children in order to fight off the moral hazard that results from their potential provision of financial support in case their children fail in the labor market. Their work ethics lead children to increase their efforts in human capital acquisition and, thereby, increase the likelihood of labor market success. This raises the question whether the model of L&N would already capture the "ethics of law abidance" if one simply reinterpreted high (low) efforts in human capital acquisition as low (high) criminal activity.

In fact, the difference between our models is not just semantic, but economic. In L&N, children's costs of effort are *sunk* when it comes to ex post transfers from parents; whereas, in the range of illegal activities, the costs of "effort" are given by the risk of punishment, which might materialize in ex post fines, and thus might affect the ex post transfers from parents. Due to this difference, our analysis is also much more involved than the one of L&N.

and institutions.[8] These studies have in common that the interaction between generations is 'simple', in particular not involving problems of intergenerational moral hazard, as it is the case in Lindbeck and Nyberg (2006) and our study. As we are particularly interested in the role of moral hazard for vertical cultural transmission, we abstract from the interaction between vertical and horizontal transmission. This allows us to determine endogenously how parenting and institutions (formal law enforcement in our case) jointly shape the various externalities that children impose on parents and how parenting induces children to (partially) internalize these externalities. To the best of our knowledge, we add to the literature on 'cultural transmission' a novel perspective: the role of expected intergenerational social mobility.

As we investigate how formal law enforcement affects the formation of people's ethics of law abidance, our study naturally relates to the literature on law and economics. Following the pioneering work of Becker (1968) and Becker and Stigler (1974), there is a large literature on the effects and efficient design of formal law enforcement. This literature typically conducts economic analyses of different law enforcement policies in a variety of criminal settings.[9] Part of this literature investigates law enforcement in the presence of social norms or people's ethics of law abidance, typically assuming, however, that those are not affected by formal law enforcement; examples are Posner (1997), Ellickson (1998), and McAdams and Rasmusen (2007).[10] Only recently, the theory of 'motivation crowding', which investigates empirically as well as theoretically how extrinsic incentives might crowd out people's intrinsic motivations (e.g., Kreps, 1997; Frey and Jegen, 2001; and Bénabou and Tirole, 2003) has made its way into the law and economics literature. Acemoglu and Jackson (2017) make a first step in this direction by investigating how formal law enforcement might "backfire" if the law to be enforced is in conflict with a community's social norms. We contribute to this new direction in the theory of law and economics by

[8] Prominent examples are Hauk and Saez-Martib (2002), Dessi (2008), Doepke and Zilibotti (2008), Tabellini (2008), Adriani and Sonderegger (2009), Bidner and Francois (2010), and Acemoglu and Jackson (2015, 2017). For a survey, see Bisin and Verdier (2010).

[9] For an extensive survey, see Polinsky and Shavell (2000).

[10] Gordon (1989), in the range of tax compliance, has provided an early analysis of how social norms of law abidance are affected by the determinants of law (in his case, by changes in tax rates, not in law enforcement). For a literature survey on 'tax morale', see Luttmer and Singhal (2014).

studying *when*, in terms of expected intergenerational social mobility, more surveillance or tougher punishment might strengthen or weaken people's ethics of law abidance.

The results of this study emphasize the importance of expected intergenerational social mobility in affecting the formation of people's ethics of law abidance. The literature on human development and social mobility (see, e.g., Heckman and Mosso, 2014, and the references discussed therein) shows that early life conditions—"including parenting"—are important in shaping multiple life skills that can have significant effects on children's economic outcomes. Using large data sets, recent empirical studies investigate how neighborhood characteristics—such as concentrated poverty, school quality, and share of two-parent families—do affect children's earnings and other outcomes (e.g., Chetty and Hendren, 2016a, 2016b). Our study contributes to this literature by articulating the hitherto neglected insight that in impoverished neighborhoods, expected intergenerational social mobility can affect parents' incentives to bring up their children as law-abiding citizens.

## III.3. The Model

Our model consists of three periods. In the first period, a parent instills an ethic of law abidance in her adolescent child.[11] This ethic determines the shame that her child incurs when being caught breaking the law. In the second period, the grown up child decides on how much to engage in illegal activities. These might yield him some monetary returns which exceed the income he could earn legally instead. With a given probability, however, he will be convicted and has to pay a fine. In the third period, the parent decides on whether to support her child with a monetary transfer.

The parent is assumed to be altruistic towards her child in that she cares about his material well-being.[12] She might therefore be willing to provide him with a transfer. As the parent cannot commit ex ante to such a transfer, this might incentivize her child to

[11] For simplicity, we focus on a single parent ('she') and a single child ('he').

[12] Within the economic literature on parenting, two different approaches are taken to model parental incentives. The approach dominant in the family economics literature, that we will take in our study, considers the material incentives of altruistic parents, whose children anticipate this altruism (e.g., Lindbeck and Nyberg, 2006). The other approach is to consider parents as imperfectly paternalistic (e.g., Tabellini, 2008), meaning that parents steer the behavior of their children because they think they know best what is good for them.

engage too much in illegal activities, since the larger the fine when caught, the higher the transfer he can expect. Effectively, the parent provides her child with a partial insurance against being convicted. In order to alleviate the moral hazard that originates from her own inability to commit to a transfer, the parent can instill an ethic of law abidance in her child.

As the model is solved through backward induction, we start by discussing the third period.

### III.3.1. Parental Transfers

The parent is assumed to equipped with a given level of wealth, $I_P > 0$, from which she can both consume and provide a transfer to her child.[13] We take $I_P$ to capture the parent's income, savings, real estate, etc., and conveniently refer to it as her *disposable income*. The size of the transfer generally depends on whether the child has been convicted or not, and is given by $r^i = I_P - c_P^i$. Here, $c_P^i$ gives the parent's consumption, and the superscript $i \in \{d, u\}$ indicates whether the child has been *detected* ($d$) or remains *undetected* ($u$). Parental transfers are assumed to be non-negative, $r^i \geq 0$, implying that the parent cannot take away income from her child.

Utility from own consumption is given by a function $u(c)$, which is assumed to be isoelastic and identical for both parent and child. We can thus write $u(c) = \frac{c^{1-\rho}}{1-\rho}$, with $\rho > 0$ denoting the coefficient of relative risk aversion. The parent derives utility from her own consumption, and she values the utility that her child derives from his consumption. Moreover, when her child is caught breaking the law, she might incur disutility due to shame. We write parental utility as

$$U_P^i = u(c_P^i) + \alpha u(c_C^i) - S_P^i, \tag{III.1}$$

where $\alpha \in (0, 1]$ denotes the degree of parental altruism and $c_C^i$ gives the child's consumption. The child's consumption equals the sum of his pre-transfer disposable income, $I_C^i$, and the transfer: $c_C^i = I_C^i + r^i$. The term $S_P^i \geq 0$ indicates the shame that the parent

[13] Parent-specific variables will be indexed by a subscript $P$, child-specific variables by a subscript $C$.

incurs if her child is convicted. In some of the law and economics literature a distinction is made between 'shame', a disutility associated with a loss of reputation, and 'guilt', a disutility incurred when breaking the law regardless of conviction (e.g., Andreoni, Erard, and Feinstein, 1998). Although both types of disutility might be relevant in reality, in this study we focus on shame, such that $S_P^u = 0$.[14]

The parent chooses the transfer $r^i$ so as to maximize $U_P^i$ subject to $c_P^i = I_P - r^i$, $c_C^i = I_C^i + r^i$, and $r^i \geq 0$. This yields the following equilibrium condition:

$$u'(c_P^i) \geq \alpha u'(c_C^i). \tag{III.2}$$

If the parent earns sufficiently more than her child, such that $u'(I_P) < \alpha u'(I_C^i)$, then (III.2) holds with equality and the transfer is *operative*.[15] In that case, the transfer is set such that the parent is indifferent on the margin between higher consumption for herself and higher consumption for her child. If the parent does not earn sufficiently more than her child, the transfer is *inoperative*: $r^i = 0$. Since utility is assumed to be isoelastic, (III.2) can be written as

$$\frac{c_C^i}{c_P^i + c_C^i} \geq \frac{1}{1 + \alpha^{-1/\rho}} = \phi. \tag{III.3}$$

Whenever the transfer is operative, parent and child consume in fixed proportions of total shared income, where the child's proportion is determined by $\phi \in (0, \frac{1}{2}]$. Substituting for $c_P^i = I_P - r^i$ and $c_C^i = I_C^i + r^i$, we obtain the equilibrium parental transfer:

$$r^i = \max\left\{0, \phi I_P - (1 - \phi) I_C^i\right\}. \tag{III.4}$$

If the transfer is operative, it increases in the child's share in total consumption, $\phi$, and thus, according to (III.3), in parental altruism, $\alpha$, and relative risk aversion, $\rho$. Intuitively, the more altruistic the parent, the more she cares for her child; the more concave utility, the larger the gains from redistribution. Moreover, the transfer is increasing in the parent's

[14] With regard to the parent, it is not clear why she would even know that her child is engaged in illegal activities as long as he is not convicted.

[15] We follow Bruce and Waldman (1990) and say transfers are *operative* when positive, and *inoperative* when zero.

and decreasing in the child's disposable income. By (III.4), the transfer is operative only if the parent's disposable income is sufficiently higher than that of her child.

Hence, if the transfer is inoperative, consumption levels of parent and child simply equal their pre-transfer disposable incomes; if the transfer is operative, consumption levels of parent and child are fixed shares of total disposable income. With (III.3), this can be summarized as follows:

$$\text{If} \quad \frac{I_C^i}{I_P + I_C^i} \geq \phi : \qquad c_C^i = I_C^i, \qquad c_P^i = I_P. \tag{III.5}$$

$$\text{If} \quad \frac{I_C^i}{I_P + I_C^i} < \phi : \qquad c_C^i = \phi(I_P + I_C^i), \qquad c_P^i = (1-\phi)(I_P + I_C^i). \tag{III.6}$$

As we discuss below, the child must pay a fine when he is convicted. Consequently, his pre-transfer disposable income is weakly higher when he remains undetected: $I_C^u \geq I_C^d$. This implies that parental support generally depends on whether the child is convicted or not. We distinguish between three possible *regimes* of parental support:

*R*1**:** In *transfer regime R*1, parental transfers are *operative* regardless of whether the child is convicted. This is the case if the child's disposable income is sufficiently low even if his illegal activities remain undetected.

*R*2**:** In *transfer regime R*2, parental transfers are operative *if and only if* the child is convicted. When convicted, the child's disposable income is relatively low so that the parent provides him with a transfer. Otherwise, his disposable income is relatively high so that the parent does not provide him with a transfer.

*R*3**:** In *transfer regime R*3, parental transfers are *inoperative* regardless of whether the child is convicted. This is the case if the child's disposable income is sufficiently high even if his illegal activities are detected and he has to pay a fine.

### III.3.2. Crime

In the second period, the child decides on how much to engage in illegal activities, while taking into account the effect on parental transfers. He has a fixed (life-)time endowment, normalized to 1, of which he spends $\tau$ on legal labor and $1-\tau$ on illegal activities. Normalizing his legal wage rate to $y > 0$, the child's legal income prospects are given by $y$. In the following, we concentrate on the amount $x = (1-\tau)y \in [0, y]$ of legal income the child chooses to forgo and refer to it as the child's *crime level*; in particular, we treat his crime level $x$ as the child's choice variable.

Let $h(1-\tau)$ be the monetary return to crime, where $h$ is a twice continuously differentiable function satisfying $h(0) = 0$, $h' > 0$, $h'' < 0$, $\lim_{z\downarrow 0} h'(z) = \infty$, and $h'(y) = 1$. The child's pre-transfer disposable income when undetected then equals $I_C^u = \tau y + h(1-\tau) = \tau y + (1-\tau)y + [h(1-\tau) - (1-\tau)y] = y + \left[h(\frac{x}{y}) - x\right] = y + f(x)$, where $f(x)$ measures the net monetary return to crime and satisfies $f(0) = 0$, $f' > 0$, $f'' < 0$, $\lim_{x\downarrow 0} f'(x) = \infty$, and $f'(y) = 0$. We further assume that $y + f(y) > I_P$, so that a criminal career allows the child to (potentially) become richer than his parent.[16]

However, law enforcement authorities will detect the child's illegal activities with an exogenously given probability $p \in (0,1)$, the *detection rate*. Denote by $P(x)$ the monetary sanction imposed on the child in case his crimes are detected, consisting of the portion of illegal income seized by the authority and an additional fine. When detected, the child's pre-transfer disposable income equals $I_C^d = y + f(x) - P(x) = y - \pi g(x)$, where we take $g$ to be a twice continuously differentiable function satisfying $g(0) = 0$, $g(y) = y$, $g' > 0$, and $g'' \geq 0$; the exogenous parameter $\pi \in (0,1)$ specifies the *fine rate*. Notice that $\pi g(x)$ specifies the monetary sanction in excess of the net return from crime.[17] We take the detection rate, $p$, and the fine rate, $\pi$, as instruments of formal law enforcement.

In addition to the utility he derives from consumption, the child incurs disutility from being convicted. His utility takes the form

$$U_C^i = u(c_C^i) - S_C^i, \tag{III.7}$$

[16] Otherwise, the scenario characterized by Lemma III.1(i) might vanish.

[17] Notice that one can write $I_C^d = (y-x) + (x + f(x)) - (f(x) + \pi g(x))$, such that $f(x) + \pi g(x)$ specifies the effective punishment on the gross monetary returns to crime, $x + f(x)$. Capturing the idea that a humane society would not sanction crime arbitrarily heavily and that the child can retain part of the fruits he reaped from crime even when being detected (for instance, because of scant evidence underlying some allegations), we assume that $f(x) + \pi g(x) < x + f(x)$ for all crime levels $x$, particularly implying that $I_C^d = (1-\pi)y > 0$ for $x = y$.

where $S_C^i$ denotes the child's shame. When undetected, he does not incur any shame: $S_C^u = 0$. When detected, the extent of shame is assumed to be proportional to the extent of punishment: $S_C^d = s_C \pi g(x)$, where the parameter $s_C \geq 0$ indicates the child's ethic of law abidance.[18]

The child chooses $x$ so as to maximize his expected utility

$$E[U_C] = (1-p)u(c_C^u) + pu(c_C^d) - ps_C \pi g(x). \tag{III.8}$$

His optimum illegal activity satisfies the following first-order condition:[19]

$$\frac{dE[U_C]}{dx} = (1-p)u'(c_C^u)\frac{dc_C^u}{dx} + pu'(c_C^d)\frac{dc_C^d}{dx} - ps_C \pi g'(x) = 0. \tag{III.9}$$

Substituting the derivatives of his consumption, $c_C^i = I_C^i + r^i$, and his state-dependent pre-transfer disposable income, $I_C^u = y + f(x)$ or $I_C^d = y - \pi g(x)$, into (III.9) yields the following equilibrium condition for the child's illegal activity:

$$(1-p)u'(c_C^u)\left(f'(x) + \frac{dr^u}{dx}\right) = pu'(c_C^d)\left(\pi g'(x) - \frac{dr^d}{dx}\right) + ps_C \pi g'(x). \tag{III.10}$$

The left-hand side of (III.10) gives the expected marginal benefits from crime. If the child remains undetected, more illegal activity yields higher returns from crime and a potential change in parental support. The right-hand side gives the expected marginal costs of crime. The first term indicates that more illegal activity leads to higher investment costs and a higher fine in case of detection as well as to a potential change in parental support. The second term indicates that more illegal activity increases the shame from being convicted.

As parental transfers are only affected by the child's illegal activity when transfers are operative, the equilibrium level of crime depends on the relevant transfer regime. If the transfer is inoperative, then $dr^i/dx = 0$; if the transfer is operative, then $dr^i/dx =$

[18]We assume that the shame from being convicted results from a loss of reputation in society. As society has implemented the punishment policy $x \to \pi g(x)$, the fine $\pi g(x)$ can be interpreted as the *objective* loss of reputation, whereas $s_C \pi g(x)$ can be interpreted as the *subjective* loss of reputation, which is determined by the child's ethic of law abidance.

[19]Notice that the model assumptions induce interior solutions to the child's optimization problem.

$-(1-\phi)dI_C^i/dx$. Substituting the latter conditions and the respective derivatives of $I_C^i$ into (III.10) yields the child's equilibrium conditions for transfer regimes $R1$, $R2$, and $R3$:

$$R1\textbf{:} \quad \phi(1-p)u'(c_C^u)f'(x) = \phi\pi p u'(c_C^d)g'(x) + ps_C\pi g'(x); \tag{III.11}$$

$$R2\textbf{:} \quad (1-p)u'(c_C^u)f'(x) = \phi\pi p u'(c_C^d)g'(x) + ps_C\pi g'(x); \tag{III.12}$$

$$R3\textbf{:} \quad (1-p)u'(c_C^u)f'(x) = \pi p u'(c_C^d)g'(x) + ps_C\pi g'(x). \tag{III.13}$$

The left-hand sides of the first-order conditions (III.11)-(III.13) indicate that the marginal benefits from crime are, *ceteris paribus*, lower in regime $R1$ than in the other two regimes. When undetected, more illegal activity yields the child a larger return and thus more consumption. In regime $R1$, this implies a smaller transfer from the parent, reducing the marginal benefits from crime relative to regimes $R2$ and $R3$: In regime $R1$, only a share $\phi$ of the returns from crime accrues to the child.

The right-hand sides of the first-order conditions (III.11)-(III.13) show that the child's marginal costs of crime are, *ceteris paribus*, smaller in regimes $R1$ and $R2$ than in regime $R3$. When detected, more illegal activity leads to a larger fine. But with parental transfers being operative in regimes $R1$ and $R2$, a larger fine implies a larger transfer: The child has to bear only a share $\phi$ of an increase in the fine.

Which regime applies depends on the extent of the child's illegal activity as well as on how his legal income prospects compare to the wealth of his parent. For relatively high legal income prospects, $y > I_P$, he might trigger transfers to be operative when being detected by exaggerating his illegal activities, thereby moving from regime $R3$ to regime $R2$. On the other hand, for relatively low legal income prospects, $y < I_P$, he might trigger transfers to be operative even in the case that he remains undetected, thereby moving from regime $R2$ to $R1$. The following Lemma establishes under which conditions one or the other of the first-order conditions (III.11)-(III.13) applies.

**Lemma III.1** *Let $\hat{y}$ such that $\hat{y} + f(\hat{y}) = \frac{\phi}{1-\phi}I_P$. Then, for each $s_C \geq 0$, there exist unique critical levels of the child's legal income prospects, $\{y_1^*, y_2^*\}$, satisfying $y_1^* \in (\hat{y}, \frac{\phi}{1-\phi}I_P)$ and $y_2^* \in (\frac{\phi}{1-\phi}I_P, \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P)$, such that the following conditions hold.*

*(i) If $y \le \hat{y}$, regime $R1$ applies for any crime level $x$.*

*(ii) For each $y \in (\hat{y}, y_1^*)$, the child chooses $x$ such that regime $R1$ applies. For each $y \in (y_1^*, \frac{\phi}{1-\phi} I_P)$, he chooses $x$ such that regime $R2$ applies.*

*(iii) If $y = \frac{\phi}{1-\phi} I_P$, transfer regime $R2$ applies for any crime level $x$.*

*(iv) For each $y \in (\frac{\phi}{1-\phi} I_P, y_2^*)$, the child chooses $x$ such that regime $R2$ applies. For each $y \in (y_2^*, \frac{1}{1-\pi}\frac{\phi}{1-\phi} I_P)$, he chooses $x$ such that regime $R3$ applies.*

*(v) If $y \ge \frac{1}{1-\pi}\frac{\phi}{1-\phi} I_P$, regime $R3$ applies for any crime level $x$.*

**Proof.** See Appendix C.1. ■

If the child's legal income prospects are low, $y < \hat{y}$, he will always receive parental support ($R1$), even if he invests all his time in illegal activity and remains undetected. If his legal income prospects are high, $y > y_3^*$, he will never receive parental support ($R3$), even if he invests all his time in illegal activity and is detected. The maximum legal income level $y = \frac{\phi}{1-\phi} I_P$ is exactly the one for which his parent would not provide him with a transfer if he fully abided by the laws, and thus certainly not if he engaged in illegal activities and remained undetected, but for which she would support him if his crimes were detected ($R2$).

At intermediate income prospects $y \in (\hat{y}, \frac{\phi}{1-\phi} I_P)$, the child faces a trade-off between low illegal activity, resulting in positive transfers even when undetected but low transfers when detected ($R1$), and high illegal activity, resulting in zero transfers when undetected but high transfers when detected ($R2$). Similarly, at higher income prospects, $y \in (\frac{\phi}{1-\phi} I_P, \frac{1}{1-\pi}\frac{\phi}{1-\phi} I_P)$, the child faces a trade-off between low illegal activity, resulting in zero transfers regardless of conviction ($R3$), and high illegal activity, again resulting in zero transfers when undetected but high transfers when detected ($R2$).

### III.3.3. Ethics Formation

In the first period, the parent anticipates her child's legal income prospects and his illegal efforts as an adult, the latter of which will be affected by her potential provision of financial transfers as well as the ethic of law abidance she instills in him during his adolescence.

Specifically, the parent determines to what extent her child will be susceptible to the shame from being convicted; that is, she decides on the parameter $s_C \geq 0$. In order to keep the analysis tractable, we assume that there arise no direct costs to the parent from instilling $s_C$ in her child. Notice that her choice of $s_C$ will nevertheless be limited due to the indirect costs that result from her parental altruism and the fact that, through $s_C$, she distorts her child's risk-taking decision. Moreover, we abstract from the indirect costs the parent might incur from compassion: Due to parental altruism, the child's shame from being convicted might negatively affect the parent's *ex post* utility. We assume that the parent does not internalize these costs when bringing up her child. Ex ante, her mere concerns when instilling an ethic of law abidance in her child are to fight off the moral hazard that might result form her partial insurance provision and to limit her own shame in case her child is convicted, while taking into account the forgone illegal income her child could raise.[20]

As in the case of the child, we assume that parental shame is proportional to the punishment her child incurs when being convicted. Specifically, we assume that $S_P^d = s_P \pi g(x)$, with an exogenous parameter $s_P \geq 0$ indicating how susceptible the parent is to the shame from being exposed as the parent of a criminal. According to (III.1), the parent's expected utility is thus given by

$$E[U_P] = (1-p)[u(c_P^u) + \alpha u(c_C^u)] + p[u(c_P^d) + \alpha u(c_C^d)] - p s_P \pi g(x). \qquad \text{(III.14)}$$

[20]It seems plausible that, qualitatively, abstracting from these direct or indirect costs does neither affect the shape of equilibrium ethic as a function of the child's legal income prospects (as depicted in Figure III.1 below) nor the comparative statics of ethics formation with respect to changes in formal law enforcement. Notice, however, that when incorporating these costs into the parent's expected utility, the absolute level of equilibrium ethics, $s_C^*$, might be much lower than depicted in Figure III.1; potentially even $s_C^* < s_P$ in transfer regime $R3$.

She chooses $s_C$ so as to maximize her expected utility, taking into account the disincentive effect on her child's illegal efforts and the respective financial support she (potentially) will provide him with later in life. Because her child's illegal efforts might change discretely in response to a marginal increase in $s_C$, as he switches from one transfer regime to another, the parent's expected utility is not everywhere continuous in $s_C$. The equilibrium value of $s_C$ could therefore either be 'interior' or not. We first focus on interior equilibria; we discuss non-interior equilibria and equilibrium selection later on.

Maximizing the parent's expected utility with respect to $s_C \geq 0$, while substituting for (III.9) and $\frac{dc_P^i}{s_C} = \frac{-dr^i}{s_C}$, yields[21]

$$\frac{dE[U_P]}{ds_C} = p(s_P - \alpha s_C)\pi g'(x)\frac{-dx}{ds_C} - pu'(c_P^d)\frac{dr^d}{ds_C} - (1-p)u'(c_P^u)\frac{dr^u}{ds_C} = 0. \qquad \text{(III.15)}$$

Instilling stronger ethics in her child affects the parent's expected utility in three different ways. First of all, as it enhances the child's law abidance, it reduces the parent's expected shame by $ps_P\pi g'(x)\frac{-dx}{ds_C} > 0$. However, this comes at the expense of the child's worsening his own material well-being. As the parent values her child's material well-being by $\alpha$, this cost to the parent can be written as $-\alpha ps_C\pi g'(x)\frac{-dx}{ds_C} < 0$. Thus, the first term in (III.15) measures the parent's trade-off between reducing her own shame when her child is convicted and distorting the child's decision to engage in illegal activities. Ignoring the other terms in (III.15), raising $s_C$ yields a benefit for the parent as long as $s_P > \alpha s_C$. The second and third effects of a stronger ethic on the parent's expected utility originate from its effects on parental transfers. As a stronger ethic leads the child to engage less in illegal activities (according to (III.11)-(III.13)), his disposable income will be higher in case of conviction, but lower otherwise. As a result, depending on the relevant transfer regime, the child might receive a lower transfer in case of conviction, and a higher transfer

[21] That is, taking the derivative of (III.14) with respect to $s_C$ and substituting for $\frac{dc_P^i}{s_C} = \frac{-dr^i}{s_C}$ yields

$$\frac{dE[U_P]}{ds_C} = (1-p)\left(-u'(c_P^u)\frac{dr^u}{ds_C} + \alpha u'(c_C^u)\frac{dc_C^u}{ds_C}\right) + p\left(-u'(c_P^d)\frac{dr^d}{ds_C} + \alpha u'(c_C^d)\frac{dc_C^d}{ds_C}\right) - ps_P\pi g'(x)\frac{dx}{ds_C}.$$

Furthermore, note that the child's consumption is only affected by $s_C$ through its effect on $x$, such that we can write $\frac{dc_C^i}{ds_C} = \frac{dc_C^i}{dx}\frac{dx}{ds_C}$. Substituting for $\frac{dc_C^i}{dx}$ from the individual's first-order condition given by (III.9) yields (III.15).

otherwise. The lower transfer in case of conviction brings a benefit for the parent, given by the second term in (III.15). The higher transfer if illegal activities remain undetected imposes a cost on the parent, given by the third term in (III.15).

Since the effects on transfers depend on the specific transfer regime, the equilibrium ethic of law abidance also depends on the transfer regime. Recall that whenever the transfer is inoperative, $\frac{dr^i}{ds_C} = \frac{dr^i}{dx}\frac{dx}{ds_C} = 0$. Whenever the transfer is operative, (III.4) implies that $\frac{dr^d}{dx} = -(1-\phi)\frac{dI_C^d}{dx} = (1-\phi)\pi g'(x)$ and $\frac{dr^u}{dx} = -(1-\phi)\frac{dI_C^u}{dx} = -(1-\phi)f'(x)$. Substituting for the change in transfers, we can rewrite (III.15) to obtain the equilibrium ethic of law abidance associated with transfer regimes $R1$, $R2$, and $R3$:[22]

$$\mathbf{R1:} \qquad \alpha s_C = \phi s_P, \tag{III.16}$$

$$\mathbf{R2:} \qquad \alpha s_C = s_P + (1-\phi)u'(c_P^d), \tag{III.17}$$

$$\mathbf{R3:} \qquad \alpha s_C = s_P, \tag{III.18}$$

where the optimal level of illegal activity is determined by the child's first-order condition in the respective transfer regime, given by (III.11)-(III.13).

The stronger the child's ethic of law abidance the less he engages in illegal activity. His ethic of law abidance distorts his trade-off between consumption in either state of the world, causing him to sacrifice expected utility from consumption in order to alleviate the expected shame from being convicted. The marginal costs for the parent of further strengthening her child's ethic of law abidance are given by the left-hand sides of (III.16)-(III.18). The extent to which the child's material trade-off is distorted is measured on the margin by $s_C$. The marginal costs to the parent of further strengthening her child's ethic are given by $\alpha s_C$. Marginal benefits of reduced illegal efforts depend on the relevant regime and are given by the right-hand sides of (III.16)-(III.18).

In regime $R1$, the parent provides her child with a transfer regardless of whether he is convicted or not. Consequently, the parent cares for her child's material well-being not only because of her altruism towards him, but also because she indirectly shares in his income. The higher the parent's share in total income (that is, the lower $\phi$), the costlier it

[22]We make use of (III.2) and (III.11) to obtain condition (III.16).

is for her to worsen her child's material well-being by distorting his incentives to earn some extra money from illegal activities. As illustrated by (III.16), in trading off her child's material well-being with her own shame, she discounts the shame by a factor $\phi$.

In regime $R2$, the parent provides her child with a transfer if and only if he is convicted. This transfer increases in his illegal effort. The parent, therefore, has a material incentive to fight off the resulting moral hazard by instilling a stronger ethic of law abidance in her child. A marginal decrease in the child's illegal effort increases his disposable income in case of conviction by $\pi g'(x)$, a share $1 - \phi$ of which accrues to the parent. Hence, the parent's marginal material benefit of less crime committed by her child is given by the second term on the right-hand side of (III.17). The total marginal benefit of further distorting the child's material trade-off are thus twofold: it reduces parental shame and parental support if the child is convicted.

In regime $R3$, the parent never provides her child with a transfer and thus has only an intangible incentive to bring up her child as a law-abiding citizen. A marginal reduction of the child's illegal efforts reduces parental shame by $s_P$. The intuition behind (III.18) is thus straightforward. When deciding on how much to engage in crime, the child does not take into account the associated shame his parent might incur. The parent forces the child to internalize this negative externality on her by instilling in him an ethic of law abidance, up to the point where her marginal costs of distorting her child's material trade-off equal her marginal benefits from lower shame. Since transfers are absent, the parent has no direct stake in her child's material well-being, so she instills in him a stronger ethic of law abidance than she would in regime $R1$.

Lemma III.1 establishes that for intermediary levels of the child's legal income prospects, regimes $R1$ and $R3$ are incentive compatible only if the child's shame from being convicted is larger than some income-dependent critical level. Thus, the values for $s_C$ that are implied by (III.16)-(III.18) might or might not be incentive compatible. If they are, we call the equilibrium level of $s_C$ *interior*. The following Lemma establishes under what conditions the equilibrium value for $s_C$ is interior, and it characterizes the equilibrium value for $s_C$ in case it is not interior.

**Proposition III.1** *There exist four critical levels of the child's legal income prospects, $\{\tilde{y}_{11}, \tilde{y}_{12}, \tilde{y}_{23}, \tilde{y}_{33}\}$, satisfying*

$$\hat{y} < \tilde{y}_{11} < \tilde{y}_{12} < \frac{\phi}{1-\phi} I_P < \tilde{y}_{23} < \tilde{y}_{33} < \frac{1}{1-\pi} \frac{\phi}{1-\phi} I_P,$$

*such that the following conditions hold.*[23]

*(i) If $y < \tilde{y}_{11}$, equilibrium ethic is given by $s_C^*(y) = \frac{\phi}{\alpha} s_P$.*

*(ii) For each $y \in (\tilde{y}_{11}, \tilde{y}_{12})$, equilibrium ethic is given by the unique $s_C^*(y)$ for which the child is indifferent between regimes $R1$ and $R2$. The child prefers $R2$ over $R1$ if $s_C < s_C^*(y)$, and vice versa. Furthermore, $\frac{ds_C^*(y)}{dy} > 0$.*

*(iii) If $y \in (\tilde{y}_{12}, \tilde{y}_{23})$, equilibrium ethic solves $s_C^*(y) = \frac{1}{\alpha} s_P + \frac{1}{\alpha}(1-\phi) u'(c_P^d)$.*

*(iv) For each $y \in (\tilde{y}_{23}, \tilde{y}_{33})$, equilibrium ethic is given by the unique $s_C^*(y)$ for which the child is indifferent between regimes $R2$ and $R3$. The child prefers $R2$ over $R3$ if $s_C < s_C^*(y)$, and vice versa. Furthermore, $\frac{ds_C^*(y)}{dy} < 0$.*

*(v) If $y > \tilde{y}_{33}$, equilibrium ethic is given by $s_C^*(y) = \frac{1}{\alpha} s_P$.*

*The so defined function $s_C^*(y)$ is continuous in $y$.*

**Proof.** See Appendix C.2. ■

The proof of Proposition III.1 builds on the following line of reasoning. For low legal income prospects (as compared to the parent's wealth), the child receives a transfer regardless of whether he is convicted or not. If the child's legal income prospects (hypothetically) increase, then, at some point ($y = \tilde{y}_{11}$), the child prefers regime $R2$ over $R1$ if the parent instills in him the ethic associated with regime $R1$, as given by (III.16). At that point, the parent still strictly prefers regime $R1$. Consequently, she instills a stronger ethic in her child in order to keep him just indifferent between regimes $R1$ and $R2$. From

[23] As defined in Lemma III.1, the legal income level $\hat{y}$ solves $\hat{y} + f(\hat{y}) = \frac{\phi}{1-\phi} I_P$.

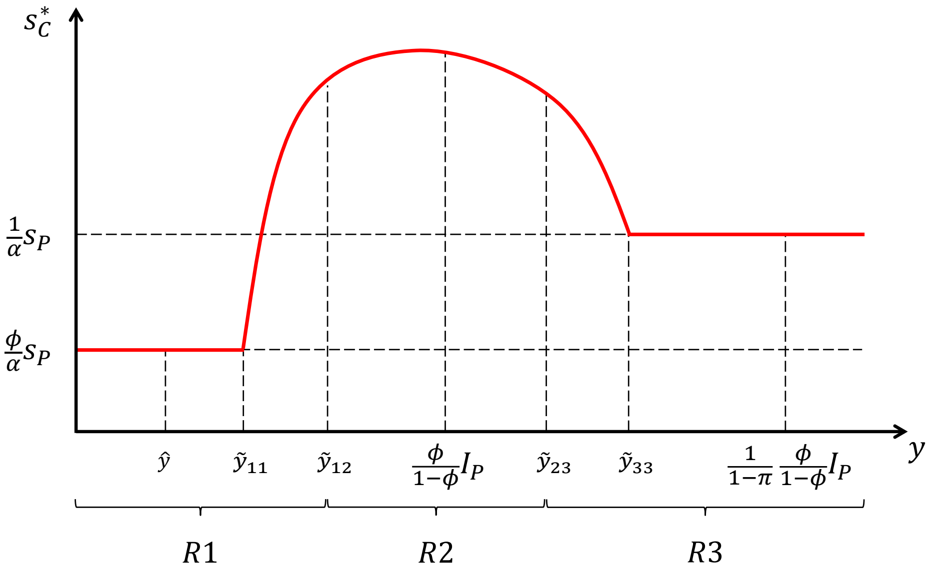


Figure III.1.: *The strength of the ethic of law abidance, $s_C^*$, that a parent instills in her child if his legal income prospects amount to $y$, while her own wealth amounts to $I_P$. In transfer regime $R2$, she supports her child financially if and only if he is convicted of illegal activities. In transfer regime $R1$ ($R3$), she does (not) support him, regardless of whether he is convicted.*

a certain level of income upwards, ($y \geq \tilde{y}_{12}$), both parent and child strictly prefer regime $R2$ with equilibrium ethic as given by (III.17).

For high legal income prospects, the child receives no transfer regardless of whether he is convicted or not, establishing transfer regime $R3$. If the child's legal income prospects (hypothetically) *decrease*, then, at some point ($y = \tilde{y}_{33}$), the child prefers regime $R2$ over $R3$ if the parent instills in him the ethic associated with regime $R3$, as given by (III.18). At that point, the parent still strictly prefers regime $R3$. Consequently, she instills a stronger ethic in her child in order to keep him just indifferent between regimes $R3$ and $R2$. From a certain level on, ($y = \tilde{y}_{23}$), even lower legal income prospects cause both parent and child to strictly prefer regime $R2$ with equilibrium ethic as given by (III.17).

Figure III.1 provides a stylized illustration of the equilibrium ethic of law abidance that the parent instills in her child. The depicted critical income prospects have been characterized by Lemma III.1 and Proposition III.1. Notice also that $\phi \leq \frac{1}{2}$, and $\frac{\phi}{1-\phi} \leq 1$.

## III.4. Intergenerational Social Mobility and Ethics Formation

As Figure III.1 illustrates, Proposition III.1 establishes a non-monotonic relationship between the legal income prospects a parent anticipates for her child (as compared to her own wealth) and the parent's incentive to instill an ethic of law abidance in him.

If her child's legal income prospects exceed her own by a sufficient amount ($y > \tilde{y}_{33}$), then the parent expects herself to never support her child financially later in life (regime $R3$). We can interpret this scenario as *intergenerational upward mobility.* In this case, the parent's mere incentive to bring up her child as a law-abiding citizen is to limit her own shame in case he is convicted of illegal activities. By Proposition III.1, transfer regime $R3$ is essentially the only relevant one if parental altruism is very weak, since $\lim_{\alpha \downarrow 0} \frac{\phi}{1-\phi} = 0$. In this case, the child's ethic is very strong, $\lim_{\alpha \downarrow 0} \frac{s_P}{\alpha} = \infty$, and results in nearly full compliance, $\lim_{\alpha \downarrow 0} x = 0$. In the following, we assume that parental altruism is sufficiently strong, such that transfer regime $R3$ applies *only if* the child's legal income prospects exceed the parent's wealth: $\frac{1}{1-\pi} \frac{\phi}{1-\phi} I_P > I_P$, which is clearly satisfied if $\alpha$ approaches 1. We take the equilibrium ethic $s_C^* = \frac{1}{\alpha} s_P$ associated with interior solutions under regime $R3$ as a reference level and refer to it as a *moderate* ethic of law abidance.

If her child's legal (as well as illegal) income prospects are sufficiently poor ($y < \tilde{y}_{11}$), be it due to a lack of capability or opportunity, then parental support will be provided even if the child succeeds in his illegal activities (regime $R1$). We might interpret this scenario as *intergenerational downward mobility.* Ex ante, the parent then benefits materially from her child's illegal activities, since she might share in his illegal income through a reduction in financial support. The countervailing incentive to still bring up her child as a law-abiding citizen is to limit her own shame in case he is convicted. Consequently, if the parent is susceptible to this sort of emotion ($s_P > 0$), the equilibrium ethic of law abidance is strictly weaker under intergenerational downward mobility than under intergenerational upward mobility, $s_C^* = \frac{\phi}{\alpha} s_P \leq \frac{1}{2} \frac{1}{\alpha} s_P$. In particular, the ethic of law abidance associated with intergenerational downward mobility might even fall short of the parent's ethic: $s_C^* < s_P$. For instance, $\frac{\phi}{\alpha} s_P < s_P$ holds if parental altruism is sufficiently strong.[24]

The equilibrium ethic of law abidance is strongest if the child's legal income prospects are 'intermediate' in the sense that the parent would provide her child with financial support if and only if he is convicted of illegal activities and suffered from a hefty fine (regime $R2$, with $y \in (\tilde{y}_{12}, \tilde{y}_{23})$). In this case, the parent's incentive to bring up a law-abiding citizen is twofold. Again, she wants to limit her own shame in case her child is convicted. In addition, she seeks to fight off the moral hazard that results from the partial insurance she provides him with. If parental altruism is strong, $\alpha \approx 1$, such that $\phi \approx \frac{1}{2}$, then transfer regime $R2$ is associated with *intergenerational stagnation*: $y \approx I_P$.

Notice that the major qualitative difference between transfer regime $R2$ on the one hand and transfer regimes $R1$ and $R3$ on the other lies in the fact that, in regime $R2$, the parent instills an ethic of law abidance in her child even if she is not susceptible to the shame from having a convicted child: $s_C^* > 0$, even if $s_P = 0$.

The above implications are independent from absolute wealth levels. They hold regardless of whether the parent belongs to the lower, middle, or upper class. However, as we have discussed in the introductory section, parents who belong to the middle or upper class might not be led in the first place by the economic incentives we have analyzed when bringing up their children as law-abiding citizens.

## III.5. Law Enforcement and Ethics Formation

How does formal law enforcement affect the equilibrium ethics of law abidance? To answer this question, we determine the equilibrium implications of increases in the detection rate and the fine rate. The following Lemma establishes the comparative statics with respect to $p$ and $\pi$ for the interior equilibria characterized by Proposition III.1(i),(iii),(v).[25]

[24] We should be cautious, however, when it comes to interpretations of the inverse relation, $s_C^* > s_P$, as we observe it in transfer regime $R3$ and parts of transfer regime $R2$. This result might not be robust to the introduction of additional 'parenting costs'. See the corresponding discussion in footnote 20.

[25] As we find (and state in Lemma III.3 below) that an increase in the fine rate has an ambiguous effect on ethics formation for non-interior equilibria, we omit the comparative statics with respect to $\pi$ for interior equilibria under regime $R2$.

**Lemma III.2** *For interior equilibria, if transfers are always operative,* $y \leq \tilde{y}_{11}$*, or never operative,* $y \geq \tilde{y}_{33}$*, the detection rate and the fine rate have no effect on ethics formation. With transfers being operative if and only if the child is convicted,* $y \in (\tilde{y}_{12}, \tilde{y}_{23})$*, an increase in the detection rate weakens the ethics of law abidance.*

**Proof.** See Appendix C.3. ■

With regard to the detection rate, the next Lemma establishes a similar crowding-out effect for the non-interior equilibria characterized by Proposition III.1(ii),(iv). It establishes in particular that an increase in the fine rate has an ambiguous effect on ethics formation.

**Lemma III.3** *For non-interior equilibria, if* $y \in (\tilde{y}_{11}, \tilde{y}_{12})$ *or* $y \in (\tilde{y}_{23}, \tilde{y}_{33})$*, an increase in the detection rate,* $p$*, weakens the ethics of law abidance. Furthermore,* $\frac{\partial \tilde{y}_{11}}{\partial p} \geq 0$ *and* $\frac{\partial \tilde{y}_{33}}{\partial p} \leq 0$*. If* $y \in (\tilde{y}_{11}, \tilde{y}_{12})$*, an increase in the fine rate strengthens the ethics of law abidance, whereas, if* $y \in (\tilde{y}_{23}, \tilde{y}_{33})$*, an increase in the fine rate can weaken the ethics of law abidance.*

**Proof.** See Appendix C.4. ■

Intuitively, the effect of an increase in the detection rate is twofold. As they become riskier, the (adult) child will reduce his illegal activities, thereby alleviating the need for parental support in case he is convicted. On the one hand, the detection rate thus acts as a substitute for ethics formation. On the other hand, depending on whether transfers are operative if the child is convicted, a higher detection rate makes a large parental transfer as well as parental shame more likely, thereby partially incentivizing the parent to instill a stronger ethic of law abidance in her child. By Lemmas III.2 and III.3, the former effect strictly dominates the latter for ‘intermediate’ legal income prospects of the child.

The effect of an increase in the fine rate is not clear-cut. On the one hand, the child will reduce his illegal activities in response to higher marginal fines. This reduction, however, might or might not result in smaller *absolute* fines in case of conviction. A higher fine rate might therefore imply a larger or smaller parental transfer, and stronger or weaker parental shame. By Lemma III.3, a higher fine rate can indeed strengthen or weaken the ethics of law abidance, depending on the child's legal income prospects as well as the strength of parental altruism.

With terminology as introduced in Section III.4, Lemmas III.2 and III.3 give proof of the following result if we assume that parental altruism is sufficiently strong, such that transfer regime $R3$ applies *only if* the child's legal income prospects exceed the parent's wealth by a sufficient amount (in terms of Proposition III.1, only if $y \geq \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P > I_P$).

**Proposition III.2** *Under strong intergenerational upward or downward mobility, changes in formal law enforcement do not affect parents' incentives to bring up their children as law-abiding citizens. Under intergenerational stagnation as well as weak upward or downward mobility, more surveillance crowds out the ethics of law abidance, whereas the effect of tougher punishment is ambiguous.*

## III.6. Conclusion

The extent to which people abide by the law is commonly believed to depend not only on formal law enforcement but also on social norms and people's ethical convictions regarding law abidance. While the related law and economics literature has largely focused on the emergence of social norms of law abidance and how those are affected by formal law enforcement, this study sheds light on a natural channel through which people's *ethics* of law abidance are presumably shaped: namely, the economic incentives of parents to bring up their children as law-abiding citizens.

We have shown that these incentives vary with the intergenerational economic standing of the family. We identify a non-monotonic relationship between expected intergenerational social mobility and the strength of the ethics of law abidance that parents instill in their children. The ethics of law abidance are weakest under intergenerational downward mobility, they are moderate under intergenerational upward mobility, and strongest under intergenerational stagnation.

The driving force behind these implications is the moral hazard that results from the partial insurance that altruistic parents (potentially) provide to their children: Depending on their children's legal income prospects, altruistic parents might expect themselves to bail out their adult children when those are convicted and punished. Their adult children might therefore engage too much in illegal activities. This, in turn, provides parents with the incentive to fight off the resulting moral hazard by instilling in their children an ethic of law abidance while they are adolescent. The incentive is strongest under intergenerational stagnation, as parents would only support their children in case of conviction and punishment. By contrast, the incentive is weakest under intergenerational downward mobility: If their children's legal (and illegal) income prospects are relatively poor, parents expect themselves to support their adult children in any case. Ex ante, they thus share in their children's illegal income through a reduction in financial support.

The effects of changes in formal law enforcement on ethics formation are sensitive to intergenerational social mobility. Under intergenerational stagnation as well as weak upward or downward mobility, a higher detection rate substitutes for parents' need to fight off the moral hazard that results from their partial insurance provision. A higher detection rate thus crowds out the ethics of law abidance. Hence, under intergenerational stagnation as well as weak upward or downward mobility, more surveillance might backfire, as it crowds out people's ethical convictions to abide by the law and, thereby, comes at double costs. We find that the effect of tougher punishment on ethics formation is ambiguous under intergenerational stagnation.

# Conclusion

In this dissertation, I have explored the role of other-regarding preferences in the form of altruism, spite, or status considerations in the resolution of adverse selection and moral hazard. In two chapters, I have reflected upon the implications of asymmetric information on other-regarding preferences for incentive mechanism design in general and human resource management in specific. In a third chapter, my co-authors, Aart Gerritsen and Vai-Lam Mui, and I have reflected upon how moral hazard within families may generate and shape people's ethical convictions regarding their abidance by social norms and formal law.

The central result of Chapter I is that the welfare judgment inherent to an allocation rule, which translates agents' privately known preferences into final allocations, is critical to whether that allocation rule can be implemented in an ex post Pareto efficient way. Implementation of a welfare judgment inconsistent with externality-ignoring utilitarianism inevitably violates budget balance and, thus, involves incentive costs.

The result has two immediate implications. In the range of conflict resolution, it provides a rationale for the 'common sense' approach most people would adopt when arbitrating between conflicting parties: namely, to *not* condition the arbitration process or the final resolution on the extent to which the opponents despise each other, but to rather "focus on the issue" and to base arbitration merely on how it would affect the opponents' *material* wealth. This is certainly the way how parents tend to resolve animosity between their children, how judges approach the resolution of divorce battles, and how third-party diplomats try to conciliate rival tribes or nations.

More generally, the result suggests that public economic policies dedicated to maximize a social welfare measure inconsistent with externality-ignoring utilitarianism do either provide people with adverse incentives (e.g., to reduce their labor supply) or are not budget-balanced, either leading to a reduction of aggregate wealth or requiring an increase of public debt.

Chapter II has shown that, if a group of coworkers forms teams autonomously in order to engage in parallel team production, then the social network of coworkers' interpersonal relationships can be such that team formation collides with overall efficient production. From a firm's perspective, this result calls for interventionist human resource management when it comes to the assignment of coworkers to teams, rather than the delegation of such responsibility to the workers.

However, in order to render team assignment decisions efficient, a staffing manager needs to have information on her staff's social network of interpersonal relationships. Nowadays, in the digital age, such information can be gathered from the 'traces' coworkers leave whenever communicating digitally. The mere existence of consultancies offering exactly this service is evidence for firms being in need of such information, materialized in their willingness to pay for it and, thus, being a matter of revealed preference. Chapter II provides a rationale for such preference.

While industrial policy, at least in Europe, is increasingly concerned with "privacy at the workplace", the central result of Chapter II indicates that such policies should not overreach. To have access to information about its coworkers' interpersonal relationships can be critical to a firm's success.

Two results are central to Chapter III. First, there exists a non-monotonic relationship between the endogenous formation of people's ethics of law abidance and the intergenerational economic standing of their families: While weak or moderate ethics result respectively from intergenerational downward or upward mobility, strong ethics result from intergenerational stagnation. And second, under strong intergenerational upward or downward mobility, a higher detection rate or tougher punishment leave ethics formation unaffected. Under intergenerational stagnation, however, a higher detection rate crowds out ethics formation, whereas the effect of tougher punishment is ambiguous.

These findings suggest that governmental efforts of formal law enforcement should condition on the (anticipated) intergenerational social mobility of people. Under intergenerational stagnation as well as weak upward or downward mobility, more surveillance might backfire, as it crowds out people's ethical convictions to abide by the law and, thereby, comes at double costs. By contrast, in phases of heavy and sustained economic downturn, potentially involving high unemployment among youth, more surveillance and tougher punishment might compensate for young people's weak ethics of law abidance.

# Appendices

# Appendix A.

# Appendix to Chapter I

## A.1. Proof of Proposition I.1

Suppose in the following that $h_i \equiv 0$ and $\Delta_i \subset [0,1)$ for all $i$. Hence, $\pi_i = \theta_i v_i$. Obviously, the social welfare measures (iv) and (v) satisfy the Pareto-property as well as condition (i) of Definition I.1. In the following, it is shown that they also satisfy the identities (I.3) and (I.4). For this purpose, ease notation by letting $\pi_i = \pi_i(k^* \,|\, \theta_i)$ and $v_i = v_i(k^*)$.

### A.1.1. Proof of Proposition I.1(iv)

Let $V(k) = [\pi_1(k \,|\, \theta_1) + \delta_1 \pi_2(k \,|\, \theta_2)] \cdot [\pi_2(k \,|\, \theta_2) + \delta_2 \pi_1(k \,|\, \theta_1)]$. By assumption, $k^* : \Theta \times \Delta \to K$ satisfies the FOC

$$0 = \frac{dV(k^*)}{dk} = \left( \frac{d\pi_1}{dk} + \delta_1 \frac{d\pi_2}{dk} \right) (\pi_2 + \delta_2 \pi_1) + \left( \frac{d\pi_2}{dk} + \delta_2 \frac{d\pi_1}{dk} \right) (\pi_1 + \delta_1 \pi_2) . \tag{A.1}$$

Define $x_1 = \pi_1 + \delta_1 \pi_2$ and $x_2 = \pi_2 + \delta_2 \pi_1$. Notice that $x_1, x_2 > 0$. Then (A.1) can be rewritten so as to obtain

$$0 = (x_1 + \delta_1 x_2) \frac{d\pi_2}{dk} + (x_2 + \delta_2 x_1) \frac{d\pi_1}{dk}, \tag{A.2}$$

where $x_1 + \delta_1 x_2 > 0$ and $x_2 + \delta_2 x_1 > 0$, since $\delta_1, \delta_2 \geq 0$. Implicit differentiation of (A.1) with respect to $\theta_1$ yields $\partial k^*/\partial\theta_1 = -X_1/[d^2V(k^*)/dk^2]$, where

$$X_1 = x_2 \frac{dv_1}{dk} + \delta_2 v_1 \left( \frac{d\pi_1}{dk} + \delta_1 \frac{d\pi_2}{dk} \right) + \delta_2 x_1 \frac{dv_1}{dk} + v_1 \left( \frac{d\pi_2}{dk} + \delta_2 \frac{d\pi_1}{dk} \right).$$

Since $d^2V(k^*)/dk^2 < 0$ by the SOC, $\mathrm{sgn}(\partial k^*/\partial\theta_1) = \mathrm{sgn}(X_1)$. Having assumed that $h_i \equiv 0$, one can make use of the identities $v_1 \frac{d\pi_1}{dk} = \pi_1 \frac{dv_1}{dk}$ and (A.2) to rewrite $X_1$ as

$$\begin{aligned} X_1 &= (x_2 + \delta_2 x_1)\frac{dv_1}{dk} + v_1 \left[ 2\delta_2 \frac{d\pi_1}{dk} + (1 + \delta_1\delta_2)\frac{d\pi_2}{dk} \right] \\ &= (x_2 + \delta_2 x_1)\frac{dv_1}{dk} + v_1 \frac{d\pi_1}{dk} \left[ 2\delta_2 - (1 + \delta_1\delta_2)\frac{(x_2 + \delta_2 x_1)}{(x_1 + \delta_1 x_2)} \right] \\ &= \frac{dv_1}{dk} \frac{Y_1}{(x_1 + \delta_1 x_2)}, \end{aligned}$$

where $Y_1 = [(x_1 + \delta_1 x_2)(x_2 + \delta_2 x_1) + \pi_1(1 - \delta_1\delta_2)(\delta_2 x_1 - x_2)]$. As $\delta_1, \delta_2 \in [0, 1)$ and $\pi_i, x_i >$ 0, letting $\delta_1 = \delta_2 = 0$ yields the lower bound $Y_1 > x_1 x_2 + \pi_1(-x_2) = (x_1 - \pi_1)x_2 = \delta_1 \pi_2 x_2 \geq 0$. Hence, $\mathrm{sgn}(\partial k^*/\partial\theta_1) = \mathrm{sgn}(X_1) = \mathrm{sgn}(dv_1/dk)$, while, by assumption, $\mathrm{sgn}(\partial k^*/\partial\theta_i) \neq 0$ and $\mathrm{sgn}(dv_i/dk) \neq 0$ for all $i$. Hence, $1 = \mathrm{sgn}^2(\partial k^*/\partial\theta_1) = \mathrm{sgn}(dv_1/dk)\,\mathrm{sgn}(\partial k^*/\partial\theta_i) =$ $\mathrm{sgn}(\partial v_1/\partial\theta_1)$. By symmetry, $1 = \mathrm{sgn}(\partial v_2/\partial\theta_2)$. Hence, $1 = \mathrm{sgn}(\partial v_1/\partial\theta_1 \cdot \partial v_2/\partial\theta_2) =$ $\mathrm{sgn}(\partial v_1/\partial\theta_2)\,\mathrm{sgn}(\partial v_2/\partial\theta_1)$, as required.

On the other hand, implicit differentiation of the FOC (A.1) with respect to $\delta_1$ yields $\partial k^*/\partial\delta_1 = -Z_1/[d^2V(k^*)/dk^2]$, where

$$Z_1 = x_2 \frac{d\pi_2}{dk} + \pi_2 \left( \frac{d\pi_2}{dk} + \delta_2 \frac{d\pi_1}{dk} \right).$$

Since $d^2V(k^*)/dk^2 < 0$ by the SOC, $\mathrm{sgn}(\partial k^*/\partial\delta_1) = \mathrm{sgn}(Z_1)$. By making use of (A.2), $Z_1$ can be written as

$$Z_1 = \frac{d\pi_2}{dk} \left[ x_2 + \pi_2 - \delta_2 \pi_2 \frac{(x_1 + \delta_1 x_2)}{(x_2 + \delta_2 x_1)} \right] = x_2 \frac{d\pi_2}{dk} \left[ 1 + \pi_2 \frac{(1 - \delta_1\delta_2)}{(x_2 + \delta_2 x_1)} \right].$$

Hence, $\operatorname{sgn}(\partial k^*/\partial\delta_1) = \operatorname{sgn}(Z_1) = \operatorname{sgn}(\pi_2/dk)$, such that

$$\operatorname{sgn}\left(\frac{\partial\pi_2}{\partial\delta_1}\right) = \operatorname{sgn}\left(\frac{d\pi_2}{dk}\right)\operatorname{sgn}\left(\frac{dk^*}{\partial\delta_1}\right) = \operatorname{sgn}^2\left(\frac{d\pi_2}{dk}\right) \in \{0,1\}.$$

By symmetry, $\operatorname{sgn}(\partial\pi_1/\partial\delta_2) = \operatorname{sgn}^2(d\pi_1/dk) \in \{0,1\}$. As $x_1 + \delta_1 x_2 > 0$ and $x_2 + \delta_2 x_1 > 0$, the FOC (A.1) implies that $d\pi_1/dk = 0$ if and only if $d\pi_2/dk = 0$. Hence, as required, $\operatorname{sgn}(\partial\pi_1/\partial\delta_2) = \operatorname{sgn}(\partial\pi_2/\partial\delta_1)$. Altogether, $W$ is sensitive. ∎

### A.1.2. Proof of Proposition I.1(v)

Let $V(k) = [[\pi_1(k\,|\,\theta_1) + \delta_1\pi_2(k\,|\,\theta_2)]^{-\rho} + [\pi_2(k\,|\,\theta_2) + \delta_2\pi_1(k\,|\,\theta_1)]^{-\rho}]^{-\frac{1}{\rho}}$, with $\rho \in (-1,\infty)\setminus\{0\}$. By assumption, $k^* : \Theta\times\Delta \to K$ satisfies the FOC

$$\begin{aligned} 0 \;=\; \frac{dV(k^*)}{dk} \;=\; \left[V(k^*)\right]^{1+\rho}\Bigg[&(\pi_1+\delta_1\pi_2)^{-\rho-1}\left(\frac{d\pi_1}{dk}+\delta_1\frac{d\pi_2}{dk}\right) \\ &+(\pi_2+\delta_2\pi_1)^{-\rho-1}\left(\frac{d\pi_2}{dk}+\delta_2\frac{d\pi_1}{dk}\right)\Bigg]. \end{aligned} \tag{A.3}$$

Define $x_1 = \pi_1 + \delta_1\pi_2$ and $x_2 = \pi_2 + \delta_2\pi_1$. Notice that $x_1, x_2 > 0$. By (A.3),

$$0 = \left(x_1^{-\rho-1} + \delta_2 x_2^{-\rho-1}\right)\frac{d\pi_1}{dk} + \left(x_2^{-\rho-1} + \delta_1 x_1^{-\rho-1}\right)\frac{d\pi_2}{dk}, \tag{A.4}$$

where $x_1^{-\rho-1} + \delta_2 x_2^{-\rho-1} > 0$ and $x_2^{-\rho-1} + \delta_1 x_1^{-\rho-1} > 0$. Implicit differentiation of (A.3) with respect to $\theta_1$ yields $\partial k^*/\partial\theta_1 = -X_1[V(k^*)]^{1+\rho}/[d^2V(k^*)/dk^2]$, where

$$\begin{aligned} X_1 \;=\; &\left(x_1^{-\rho-1} + \delta_2 x_2^{-\rho-1}\right)\frac{dv_1}{dk} - (1+\rho)x_1^{-\rho-2}v_1\left(\frac{d\pi_1}{dk} + \delta_1\frac{d\pi_2}{dk}\right) \\ &-(1+\rho)x_2^{-\rho-2}\delta_2 v_1\left(\frac{d\pi_2}{dk} + \delta_2\frac{d\pi_1}{dk}\right). \end{aligned}$$

Since $d^2V(k^*)/dk^2 < 0$ by the SOC, $\text{sgn}(\partial k^*/\partial\theta_1) = \text{sgn}(X_1)$. Having assumed that $h_i \equiv 0$, one can make use of the identities $v_1\frac{d\pi_1}{dk} = \pi_1\frac{dv_1}{dk}$ and (A.4) to rewrite $X_1$ as

$$\begin{aligned} X_1 &= \frac{dv_1}{dk}\frac{Y_1}{x_2^{-\rho-1} + \delta_1 x_1^{-\rho-1}}, \quad \text{where} \\ Y_1 &= \left(x_1^{-\rho-1} + \delta_2 x_2^{-\rho-1}\right)\left(x_2^{-\rho-1} + \delta_1 x_1^{-\rho-1}\right) \\ &\quad + (1+\rho)(1-\delta_1\delta_2)x_1^{-\rho-2}x_2^{-\rho-2}(\delta_2 x_1 - x_2)\pi_1. \end{aligned}$$

Hence, $\text{sgn}(\partial k^*/\partial\theta_1) = \text{sgn}(dv_1/dk)\,\text{sgn}(Y_1)$. Similarly, when exchanging the roles of 1 and 2, one obtains $\text{sgn}(\partial k^*/\partial\theta_2) = \text{sgn}(dv_2/dk)\,\text{sgn}(Y_2)$, where $Y_2$ is defined as

$$\begin{aligned} Y_2 &= \left(x_1^{-\rho-1} + \delta_2 x_2^{-\rho-1}\right)\left(x_2^{-\rho-1} + \delta_1 x_1^{-\rho-1}\right) \\ &\quad + (1+\rho)(1-\delta_1\delta_2)x_1^{-\rho-2}x_2^{-\rho-2}(\delta_1 x_2 - x_1)\pi_2. \end{aligned}$$

Since $(\delta_2 x_1 - x_2)\pi_1 = -(1-\delta_1\delta_2)\pi_1\pi_2 = (\delta_1 x_2 - x_1)\pi_2$, one observes that $Y_1 = Y_2$.[1] Hence, as required,

$$\begin{aligned} 1 &= \text{sgn}^2\left(\frac{\partial k^*}{\partial\theta_1}\right)\text{sgn}^2\left(\frac{\partial k^*}{\partial\theta_2}\right) \\ &= \text{sgn}\left(\frac{\partial k^*}{\partial\theta_1}\right)\text{sgn}\left(\frac{dv_1}{dk}\right)\text{sgn}\left(Y_1\right)\text{sgn}\left(\frac{\partial k^*}{\partial\theta_2}\right)\text{sgn}\left(\frac{dv_2}{dk}\right)\text{sgn}\left(Y_2\right) \\ &= \text{sgn}\left(\frac{\partial v_1}{\partial\theta_2}\right)\text{sgn}\left(\frac{\partial v_2}{\partial\theta_1}\right), \end{aligned} \tag{A.5}$$

where the first equality of (A.5) holds due to the assumption that $\partial k^*/\partial\theta i \neq 0$ for all $i$. On the other hand, implicit differentiation of the FOC (A.3) with respect to $\delta_1$ yields $\partial k^*/\partial\delta_1 = -Z_1[V(k^*)]^{1+\rho}/[d^2V(k^*)/dk^2]$, where

$$Z_1 = x_1^{-\rho-2}\left[x_1\frac{d\pi_2}{dk} - \pi_2(1+\rho)\left(\frac{d\pi_1}{dk} + \delta_1\frac{d\pi_2}{dk}\right)\right].$$

[1] For $\delta_1 = \delta_2 = 0$, one observes that $Y_i = -\rho(\pi_1\pi_2)^{-\rho-1}$. Hence, $\text{sgn}(\partial k^*/\partial\theta_i) = -\text{sgn}(dv_i/dk)$ for $\rho > 0$. For this reason, I let Definition I.1 require the weaker property of $\text{sgn}(\partial v_1/\partial\theta_2)\,\text{sgn}(\partial v_2/\partial\theta_1) = 1$, instead of $\text{sgn}(\partial k^*/\partial\theta_i)\,\text{sgn}(dv_i/dk) = 1$ for all $i$.

Since $d^2V(k^*)/dk^2 < 0$ by the SOC, $\text{sgn}(\partial k^*/\partial\delta_1) = \text{sgn}(Z_1)$. By making use of (A.4), $Z_1$ can be written as

$$\begin{aligned} Z_1 &= x_1^{-\rho-2}\frac{d\pi_2}{dk}\left[x_1 - \pi_2(1+\rho)\left(\delta_1 - \frac{x_2^{-\rho-1}+\delta_1 x_1^{-\rho-1}}{x_1^{-\rho-1}+\delta_2 x_2^{-\rho-1}}\right)\right] \\ &= x_1^{-\rho-2}\frac{d\pi_2}{dk}\left[x_1 + \pi_2(1+\rho)(1-\delta_1\delta_2)\frac{x_2^{-\rho-1}}{x_1^{-\rho-1}+\delta_2 x_2^{-\rho-1}}\right]. \end{aligned}$$

Hence, $\text{sgn}(\partial k^*/\partial\delta_1) = \text{sgn}(Z_1) = \text{sgn}(\pi_2/dk)$, such that

$$\text{sgn}\left(\frac{\partial\pi_2}{\partial\delta_1}\right) = \text{sgn}\left(\frac{d\pi_2}{dk}\right)\text{sgn}\left(\frac{dk^*}{\partial\delta_1}\right) = \text{sgn}^2\left(\frac{d\pi_2}{dk}\right) \in \{0,1\}.$$

By symmetry, $\text{sgn}(\partial\pi_1/\partial\delta_2) = \text{sgn}^2(d\pi_1/dk) \in \{0,1\}$. Since $x_1^{-\rho-1} + \delta_2 x_2^{-\rho-1} > 0$ and $x_2^{-\rho-1} + \delta_1 x_1^{-\rho-1} > 0$, identity (A.4) implies that $d\pi_1/dk = 0$ if and only if $d\pi_2/dk = 0$. Hence, as required, $\text{sgn}(\partial\pi_1/\partial\delta_2) = \text{sgn}(\partial\pi_2/\partial\delta_1)$. Altogether, $W$ is sensitive. ■

## A.2. Proof of Lemma I.1

Suppose the sensitive allocation rule $k^* : \Theta \times \Delta \to \mathbb{R}$ is strongly Bayesian implemented by the ex post budget-balanced transfer scheme $T = (t_1, t_2) : \Theta \times \Delta \to \mathbb{R}^2$. Define

$$\bar{v}_i(\hat\theta_i,\hat\delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^*(\hat\theta_i,\hat\delta_i,\theta_{-i},\delta_{-i}))\big], \tag{A.6}$$

$$\bar{h}_i(\hat\theta_i,\hat\delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[h_i(k^*(\hat\theta_i,\hat\delta_i,\theta_{-i},\delta_{-i}))\big], \tag{A.7}$$

$$\bar{\pi}_{-i}(\hat\theta_i,\hat\delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_{-i}(k^*(\hat\theta_i,\hat\delta_i,\theta_{-i},\delta_{-i})\,|\,\theta_{-i})\big], \tag{A.8}$$

$$\bar{t}_i(\hat\theta_i,\hat\delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_i(\hat\theta_i,\hat\delta_i,\theta_{-i},\delta_{-i})\big], \tag{A.9}$$

$$\bar{t}_{-i}(\hat\theta_i,\hat\delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_{-i}(\hat\theta_i,\hat\delta_i,\theta_{-i},\delta_{-i})\big], \tag{A.10}$$

where $\pi_i(k\,|\,\theta_i) = \theta_i v_i(k) + h_i(k)$. For $i \in \{1,2\}$, denote by $U_i(\hat\theta_i,\hat\delta_i\,|\,\theta_i,\delta_i)$ agent $i$'s interim expected utility from reporting $(\hat\theta_i,\hat\delta_i)$ if her true type is $(\theta_i,\delta_i)$ and if agent $-i$ reports her type truthfully:

$$U_i(\hat\theta_i,\hat\delta_i\,|\,\theta_i,\delta_i) = \theta_i\bar{v}_i(\hat\theta_i,\hat\delta_i) + \bar{h}_i(\hat\theta_i,\hat\delta_i) + \bar{t}_i(\hat\theta_i,\hat\delta_i) + \delta_i\bar{\pi}_{-i}(\hat\theta_i,\hat\delta_i) + \delta_i\bar{t}_{-i}(\hat\theta_i,\hat\delta_i). \tag{A.11}$$

Ease notation by also defining $U_i(\theta_i, \delta_i) = U_i(\theta_i, \delta_i \,|\, \theta_i, \delta_i)$. Then the following must hold for all $\theta_i, \hat{\theta}_i \in \Theta_i$ and all $\delta_i, \hat{\delta}_i \in \Delta_i$:

$$
\begin{aligned}
U_i(\theta_i, \delta_i) &\geq U_i(\hat{\theta}_i, \delta_i \,|\, \theta_i, \delta_i) && \text{(A.12)} \\
&= U_i(\hat{\theta}_i, \delta_i) + (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i, \delta_i), \\
U_i(\hat{\theta}_i, \delta_i) &\geq U_i(\theta_i, \delta_i \,|\, \hat{\theta}_i, \delta_i) && \text{(A.13)} \\
&= U_i(\theta_i, \delta_i) + (\hat{\theta}_i - \theta_i)\bar{v}_i(\theta_i, \delta_i), \\
U_i(\theta_i, \delta_i) &\geq U_i(\theta_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) && \text{(A.14)} \\
&= U_i(\theta_i, \hat{\delta}_i) + (\delta_i - \hat{\delta}_i)\Big[\bar{\pi}_{-i}(\theta_i, \hat{\delta}_i) + \bar{t}_{-i}(\theta_i, \hat{\delta}_i)\Big], \\
U_i(\theta_i, \hat{\delta}_i) &\geq U_i(\theta_i, \delta_i \,|\, \theta_i, \hat{\delta}_i) && \text{(A.15)} \\
&= U_i(\theta_i, \delta_i) + (\hat{\delta}_i - \delta_i)\Big[\bar{\pi}_{-i}(\theta_i, \delta_i) + \bar{t}_{-i}(\theta_i, \delta_i)\Big].
\end{aligned}
$$

Without loss of generality, suppose $\hat{\theta}_i > \theta_i$. Then (A.12) and (A.13) imply that

$$\bar{v}_i(\hat{\theta}_i, \delta_i) \geq \frac{U_i(\hat{\theta}_i, \delta_i) - U_i(\theta_i, \delta_i)}{\hat{\theta}_i - \theta_i} \geq \bar{v}_i(\theta_i, \delta_i). \tag{A.16}$$

As $\bar{v}_i$ is continuous on $\Theta_i$, letting $\hat{\theta}_i \downarrow \theta_i$ implies that $\partial U_i(\theta_i, \delta_i)/\partial \theta_i = \bar{v}_i(\theta_i, \delta_i)$. Integrating the latter with respect to $\theta_i$ yields the identity

$$U_i(\theta_i, \delta_i) = p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds, \tag{A.17}$$

with some function $p_i : \Delta_i \to \mathbb{R}$. Similarly, suppose $\hat{\delta}_i > \delta_i$. Then (A.14) and (A.15) imply that

$$\bar{\pi}_{-i}(\theta_i, \hat{\delta}_i) + \bar{t}_{-i}(\theta_i, \hat{\delta}_i) \geq \frac{U_i(\theta_i, \hat{\delta}_i) - U_i(\theta_i, \delta_i)}{\hat{\delta}_i - \delta_i} \geq \bar{\pi}_{-i}(\theta_i, \delta_i) + \bar{t}_{-i}(\theta_i, \delta_i). \tag{A.18}$$

As $\bar{\pi}_i$ and $\bar{t}_{-i}$ are continuous on $\Delta_i$ by assumption, letting $\hat{\delta}_i \downarrow \delta_i$ implies that

$$\frac{\partial U_i(\theta_i, \delta_i)}{\partial \delta_i} = \bar{\pi}_{-i}(\theta_i, \delta_i) + \bar{t}_{-i}(\theta_i, \delta_i). \tag{A.19}$$

Integrating with respect to $\delta_i$ in (A.19) yields the identity

$$U_i(\theta_i, \delta_i) = q_i(\theta_i) + \int_{\delta_i^{\min}}^{\delta_i} \bar{\pi}_{-i}(\theta_i, r)\, dr + \int_{\delta_i^{\min}}^{\delta_i} \bar{t}_{-i}(\theta_i, r)\, dr, \tag{A.20}$$

with some function $q_i : \Theta_i \to \mathbb{R}$. Identity (A.20) and the assumptions on the functions $v_i$ imply that $U_i(\theta_i, \delta_i)$ and, thus, $p_i$ from (A.17) must be differentiable. Jointly, identities (A.17) and (A.20) imply that

$$\int_{\delta_i^{\min}}^{\delta_i} \bar{t}_{-i}(\theta_i, r)\, dr = p_i(\delta_i) - q_i(\theta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds - \int_{\delta_i^{\min}}^{\delta_i} \bar{\pi}_{-i}(\theta_i, r)\, dr. \tag{A.21}$$

Differentiating (A.21) with respect to $\delta_i$ yields

$$\bar{t}_{-i}(\theta_i, \delta_i) = \frac{\partial p_i(\delta_i)}{\partial \delta_i} - \bar{\pi}_{-i}(\theta_i, \delta_i) + \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds. \tag{A.22}$$

Ex post budget balance requires in particular that $\bar{t}_i(\theta_i, \delta_i) = -\bar{t}_{-i}(\theta_i, \delta_i)$ on $\Theta_i \times \Delta_i$, so that truthful revelation of $(\theta_i, \delta_i)$ is Bayesian incentive-compatible for agent $i$ only if $\theta_i$ satisfies the FOC

$$\begin{aligned}
0 &= \frac{\partial}{\partial \hat{\theta}_i}\Big[\theta_i \bar{v}_i(\hat{\theta}_i, \delta_i) + \bar{h}_i(\hat{\theta}_i, \delta_i) + \delta_i \bar{\pi}_{-i}(\hat{\theta}_i, \delta_i) - (1-\delta_i)\bar{t}_{-i}(\hat{\theta}_i, \delta_i)\Big]\Bigg|_{\hat{\theta}_i = \theta_i} \\
&= \theta_i \frac{\bar{v}_i(\theta_i, \delta_i)}{\partial \theta_i} + \frac{\bar{h}_i(\theta_i, \delta_i)}{\partial \theta_i} + \delta_i \frac{\bar{\pi}_{-i}(\theta_i, \delta_i)}{\partial \theta_i} - (1-\delta_i)\left[\frac{\bar{v}_i(\theta_i, \delta_i)}{\partial \delta_i} - \frac{\bar{\pi}_{-i}(\theta_i, \delta_i)}{\partial \theta_i}\right] \\
&= \theta_i \frac{\bar{v}_i(\theta_i, \delta_i)}{\partial \theta_i} + \frac{\bar{h}_i(\theta_i, \delta_i)}{\partial \theta_i} + \frac{\bar{\pi}_{-i}(\theta_i, \delta_i)}{\partial \theta_i} - (1-\delta_i)\frac{\bar{v}_i(\theta_i, \delta_i)}{\partial \delta_i} \\
&= \mathbb{E}_{\theta_{-i}, \delta_{-i}}\left[\frac{d\pi_i(k^*(\theta, \delta)\,|\,\theta_i)}{dk}\frac{\partial k^*}{\partial \theta_i} + \frac{d\pi_{-i}(k^*(\theta, \delta)\,|\,\theta_{-i}))}{dk}\frac{\partial k^*}{\partial \theta_i} - (1-\delta_i)\frac{v_i(k^*(\theta, \delta))}{\partial \delta_i}\right],
\end{aligned} \tag{A.23}$$

where the second equality is implied by identity (A.22), and where the Leibniz integral rule has been used to obtain the second and the last equality.

In order to be Bayesian implementable with a budget-balanced mechanism, $k^*$ must satisfy identity (A.23) irrespective of the specific form that the transfer scheme might take. As $k^*$ is also assumed to be *strongly* Bayesian implementable (in the manner of

Definition I.2), identity (A.23) must hold for *any* set of (non-degenerate) type distributions $\{F_{-i}, G_{-i}\}$. However, due to the assumptions on the functions $v_i$, $h_i$, and $k^*$, the argument of $\mathbb{E}_{\theta_{-i},\delta_{-i}}[\cdot]$ in (A.23) is continuous in $(\theta_{-i}, \delta_{-i})$. Hence, $k^*$ must satisfy

$$0 = \frac{d\pi_i(k^*(\theta,\delta)\,|\,\theta_i)}{dk}\frac{\partial k^*}{\partial\theta_i} + \frac{d\pi_{-i}(k^*(\theta,\delta)\,|\,\theta_{-i}))}{dk}\frac{\partial k^*}{\partial\theta_i} - (1-\delta_i)\frac{v_i(k^*(\theta,\delta))}{\partial\delta_i}$$

for all $(\theta, \delta) \in \Theta \times \Delta$. This proves the first part of Lemma I.1. For the second part, reconsider identities (A.17) and (A.22). Under truthful revelation, they jointly imply that

$$\begin{aligned}
p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s,\delta_i)\,ds \;&=\; U_i(\theta_i,\delta_i) \qquad\qquad \text{(A.24)}\\
&=\; \theta_i\bar{v}_i(\theta_i,\delta_i) + \bar{h}_i(\theta_i,\delta_i) + \bar{t}_i(\theta_i,\delta_i)\\
&\quad\; + \delta_i\bar{\pi}_{-i}(\theta_i,\delta_i) + \delta_i\bar{t}_{-i}(\theta_i,\delta_i)\\
&=\; \theta_i\bar{v}_i(\theta_i,\delta_i) + \bar{h}_i(\theta_i,\delta_i) + \delta_i\bar{\pi}_{-i}(\theta_i,\delta_i) + \bar{t}_i(\theta_i,\delta_i)\\
&\quad\; +\delta_i\frac{\partial p_i(\delta_i)}{\partial\delta_i} - \delta_i\bar{\pi}_{-i}(\theta_i,\delta_i) + \delta_i\frac{\partial}{\partial\delta_i}\int_{\theta_i^{\min}}^{\theta_i}\bar{v}_i(s,\delta_i)\,ds\\
&=\; \theta_i\bar{v}_i(\theta_i,\delta_i) + \bar{h}_i(\theta_i,\delta_i) + \bar{t}_i(\theta_i,\delta_i)\\
&\quad\; +\delta_i\frac{\partial p_i(\delta_i)}{\partial\delta_i} + \delta_i\frac{\partial}{\partial\delta_i}\int_{\theta_i^{\min}}^{\theta_i}\bar{v}_i(s,\delta_i)\,ds.
\end{aligned}$$

Now suppose $k^*$ is independent from externality types: $k^* = k^*|_\Theta$. According to identities (A.24) and (A.22), respectively, $\bar{t}_i(\theta_i,\delta_i)$ and $\bar{t}_{-i}(\theta_i,\delta_i)$ then satisfy

$$\bar{t}_i(\theta_i,\delta_i) \;=\; p_i(\delta_i) - \delta_i\frac{\partial p_i(\delta_i)}{\partial\delta_i} - \theta_i\bar{v}_i(\theta_i,\delta_i) - \bar{h}_i(\theta_i,\delta_i) + \int_{\theta_i^{\min}}^{\theta_i}\bar{v}_i(s,\delta_i)\,ds, \quad \text{(A.25)}$$

$$\bar{t}_{-i}(\theta_i,\delta_i) \;=\; \frac{\partial p_i(\delta_i)}{\partial\delta_i} - \bar{\pi}_{-i}(\theta_i,\delta_i), \qquad \text{(A.26)}$$

where, due to $k^* = k^*|_\Theta$, only the terms containing $p_i$ effectively depend on $\delta_i$. Due to budget balance, identities (A.25) and (A.26) imply that $p_i$ solves the differential equation

$$a_i = p_i(\delta_i) + (1-\delta_i)\frac{\partial p_i(\delta_i)}{\partial\delta_i}, \qquad \text{(A.27)}$$

where $a_i$ is some constant. Differentiating (A.27) with respect to $\delta_i$ yields $\frac{\partial^2 p_i(\delta_i)}{\partial \delta_i^2} = 0$, such that $\frac{\partial p_i(\delta_i)}{\partial \delta_i} = -\alpha_i$ for some constant $\alpha_i$. Hence, identity (A.26) reads $\bar{t}_{-i}(\theta_i, \delta_i) = -\alpha_i - \bar{\pi}_{-i}(\theta_i, \delta_i)$, implying that $\bar{t}_i(\theta_i, \delta_i) = \alpha_i + \bar{\pi}_{-i}(\theta_i, \delta_i)$ due to budget balance. ■

## A.3. Proof of Theorem I.2 Continued

With notation adopted from the proof of Lemma I.1, $T^*$ satisfies

$$\begin{aligned}
\bar{t}_i(\hat{\theta}_i, \hat{\delta}_i) &= a_i + p_i(\hat{\delta}_i) - \hat{\delta}_i \frac{\partial p_i(\hat{\delta}_i)}{\partial \hat{\delta}_i} - \bar{\pi}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds - \hat{\delta}_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds, \\
\bar{t}_{-i}(\hat{\theta}_i, \hat{\delta}_i) &= b_i + \frac{\partial p_i(\hat{\delta}_i)}{\partial \hat{\delta}_i} - \bar{\pi}_{-i}(\hat{\theta}_i, \hat{\delta}_i) + \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds,
\end{aligned}$$

with appropriate constants $a_i, b_i \in \mathbb{R}$. Suppose agent $-i$ reports her type truthfully. From reporting some type $(\hat{\theta}_i, \hat{\delta}_i)$, agent $i$ of true type $(\theta_i, \delta_i)$ gains interim expected utility

$$\begin{aligned}
U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) &= \theta_i \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + \bar{h}_i(\hat{\theta}_i, \hat{\delta}_i) + \bar{t}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + \delta_i \bar{\pi}_{-i}(\hat{\theta}_i, \hat{\delta}_i) + \delta_i \bar{t}_{-i}(\hat{\theta}_i, \hat{\delta}_i) \\
&= \theta_i \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + a_i + p_i(\hat{\delta}_i) - \hat{\delta}_i \frac{\partial p_i(\hat{\delta}_i)}{\partial \hat{\delta}_i} - \hat{\theta}_i \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds - \hat{\delta}_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds \\
&\quad + \delta_i b_i + \delta_i \frac{\partial p_i(\hat{\delta}_i)}{\partial \hat{\delta}_i} + \delta_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds.
\end{aligned}$$

Partial derivatives thus satisfy

$$\frac{\partial}{\partial \hat{\theta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = (\theta_i - \hat{\theta}_i) \frac{\partial}{\partial \hat{\theta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + (\delta_i - \hat{\delta}_i) \frac{\partial}{\partial \hat{\delta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i), \tag{A.28}$$

$$\begin{aligned}
\frac{\partial}{\partial \hat{\delta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) &= (\theta_i - \hat{\theta}_i) \frac{\partial}{\partial \hat{\delta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + (\delta_i - \hat{\delta}_i) \frac{\partial^2}{\partial \hat{\delta}_i^2} \left[ p_i(\hat{\delta}_i) + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds \right].
\end{aligned} \tag{A.29}$$

Ease notation by defining $A_i = \frac{\partial}{\partial \hat{\delta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i)$, $B_i = \frac{\partial}{\partial \hat{\theta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i)$, and

$$C_i = \frac{\partial^2}{\partial \hat{\delta}_i^2} \left[ p_i(\hat{\delta}_i) + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds \right].$$

Then the partial derivatives (A.28) and (A.29) read[2]

$$\frac{\partial}{\partial \hat{\theta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = (\theta_i - \hat{\theta}_i) B_i + (\delta_i - \hat{\delta}_i) A_i, \tag{A.30}$$

$$\frac{\partial}{\partial \hat{\delta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = (\theta_i - \hat{\theta}_i) A_i + (\delta_i - \hat{\delta}_i) C_i. \tag{A.31}$$

Suppose $k^* \neq k^*|_\Theta$. Then, $B_i > 0$ by assumption. Choose $p_i(\delta_i) = \frac{1}{2} c_i \delta_i^2$, with $c_i$ as defined in (I.27). Then $C_i > 0$, and condition (I.26) is satisfied:

$$A_i^2 < B_i C_i. \tag{A.32}$$

Notice first that $(\hat{\theta}_i, \hat{\delta}_i) = (\theta_i, \delta_i)$ is the unique stationary point of $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$, as $\frac{\partial}{\partial \hat{\theta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = 0 = \frac{\partial}{\partial \hat{\delta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ implies that $(\theta_i - \hat{\theta}_i) = -(\delta_i - \hat{\delta}_i) \frac{A_i}{B_i}$ and, thus, $0 = (\delta_i - \hat{\delta}_i) \frac{1}{B_i} (B_i C_i - A_i^2)$, where $B_i > 0$ and $B_i C_i - A_i^2 > 0$. Evaluating the Hessian $H_i$ of $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ at $(\hat{\theta}_i, \hat{\delta}_i) = (\theta_i, \delta_i)$ yields

$$H_i = \begin{pmatrix} -B_i & -A_i \\ -A_i & -C_i \end{pmatrix}. \tag{A.33}$$

The principal minors of (A.33), namely $-B_i < 0$ and $\det(H_i) = B_i C_i - A_i^2 > 0$, are alternating in sign, with the first-order principal minor being negative. Hence, $(\theta_i, \delta_i)$ is a local maximizer of $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$. It remains to show that truth-telling is indeed the *unique global* expected utility maximizer for agent $i$. Given the above, it suffices to show that no local maximizer of $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ is located on the boundary of $\Theta_i \times \Delta_i$.

[2]Suppose $k^* = k^*|_\Theta$, and assume $B_i \geq 0$. Then, $A_i = 0$. When choosing $p_i = 0$, then also $C_i = 0$. By (A.30) and (A.31), truth-telling is then a global maximizer of each agent $i$'s expected utility under the transfer scheme $T^*$, which gives proof of the second part of Theorem I.2.

Suppose a local maximizer is located on $(\theta_i^{\min}, \theta_i^{\max}) \times \{\delta_i^{\min}\}$ or $(\theta_i^{\min}, \theta_i^{\max}) \times \{\delta_i^{\max}\}$. As $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ is twice partially continuously differentiable, this maximizer, $(\hat{\theta}_i, \hat{\delta}_i)$, must satisfy $0 = \frac{\partial}{\partial \hat{\theta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ and, thus, $(\theta_i - \hat{\theta}_i) = -(\delta_i - \hat{\delta}_i)\frac{A_i}{B_i}$. Substituting the latter into (A.31) yields $\frac{\partial}{\partial \hat{\delta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = (\delta_i - \hat{\delta}_i)\frac{1}{B_i}(B_i C_i - A_i^2)$. As $\frac{1}{B_i}(B_i C_i - A_i^2) > 0$, the reporting of $\hat{\delta}_i \in \{\delta_i^{\min}, \delta_i^{\max}\}$ is not optimal, which contradicts the assumption. By a similar argument one can show that no local maximizer is located on $\{\theta_i^{\min}\} \times (\delta_i^{\min}, \delta_i^{\max})$ or $\{\theta_i^{\max}\} \times (\delta_i^{\min}, \delta_i^{\max})$. Hence, only the "corners" of $\Theta_i \times \Delta_i$ qualify as potential further local maximizers.

Suppose $(\theta_i^{\max}, \delta_i^{\max})$ is a local maximizer. Then $0 \le \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\max}, \delta_i^{\max} \,|\, \theta_i, \delta_i)$ and $0 \le \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\max}, \delta_i^{\max} \,|\, \theta_i, \delta_i)$ must hold. As $(\theta_i - \theta_i^{\max}), (\delta_i - \delta_i^{\max}) < 0$, while $B_i, C_i > 0$, this implies that $A_i < 0$. However, it also implies that $(\delta_i - \delta_i^{\max}) \ge -(\theta_i - \theta_i^{\max})\frac{A_i}{C_i}$ and, thus,

$$0 \le (\theta_i - \theta_i^{\max})B_i + (\delta_i - \delta_i^{\max})A_i \le (\theta_i - \theta_i^{\max})\frac{1}{C_i}(B_i C_i - A_i^2) < 0. \tag{A.34}$$

Suppose $(\theta_i^{\max}, \delta_i^{\min})$ is a local maximizer. Then $0 \le \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\max}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ and $0 \ge \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\max}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ must hold. As $(\theta_i - \theta_i^{\max}) < 0$, while $(\delta_i - \delta_i^{\min}), B_i, C_i > 0$, this implies that $A_i > 0$. However, it also implies that $(\theta_i - \theta_i^{\max}) \ge -(\delta_i - \delta_i^{\min})\frac{A_i}{B_i}$ and, thus,

$$0 \ge (\theta_i - \theta_i^{\max})A_i + (\delta_i - \delta_i^{\min})C_i \ge (\delta_i - \delta_i^{\min})\frac{1}{B_i}(B_i C_i - A_i^2) > 0. \tag{A.35}$$

Suppose $(\theta_i^{\min}, \delta_i^{\min})$ is a local maximizer. Then $0 \ge \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\min}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ and $0 \ge \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\min}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ must hold. As $(\theta_i - \theta^{\min}), (\delta_i - \delta^{\min}), B_i, C_i > 0$, this implies that $A_i < 0$. However, it also implies that $(\delta_i - \delta^{\min}) \le -(\theta_i - \theta^{\min})\frac{A_i}{C_i}$ and, thus,

$$0 \ge (\theta_i - \theta_i^{\min})B_i + (\delta_i - \delta_i^{\min})A_i \ge (\theta_i - \theta_i^{\min})\frac{1}{C_i}(B_i C_i - A_i^2) > 0. \tag{A.36}$$

Finally, suppose $(\theta_i^{\min}, \delta_i^{\max})$ is a local maximizer. Then $0 \ge \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\min}, \delta_i^{\max} \,|\, \theta_i, \delta_i)$ and $0 \le \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\min}, \delta_i^{\max} \,|\, \theta_i, \delta_i)$ must hold. As $(\delta_i - \delta_i^{\max}) < 0$ and $(\theta_i - \theta_i^{\min}), B_i, C_i > 0$, this implies that $A_i > 0$. However, it also implies that $(\theta_i - \theta_i^{\min}) \le -(\delta_i - \delta_i^{\max})\frac{A_i}{B_i}$ and, thus,

$$0 \le (\theta_i - \theta_i^{\min})A_i + (\delta_i - \delta_i^{\max})C_i \le (\delta_i - \delta_i^{\max})\frac{1}{B_i}(B_i C_i - A_i^2) < 0. \tag{A.37}$$

Altogether, $(\theta_i, \delta_i)$ is the unique global maximizer of $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$. As the above arguments hold for any set of type distributions, $T^*$ *strongly* Bayesian implements $k^*$. ■

## A.4. Derivation of the transfer scheme $T^*$ in the proof of Theorem I.2

Suppose the transfer scheme $T^* = (t_1^*, t_2^*) : \Theta \times \Delta \to \mathbb{R}^2$ (strongly) Bayesian implements the twice continuously differentiable allocation rule $k^* : \Theta \times \Delta \to K$. With notation adopted from the proof of Lemma I.1, condition (A.22) of that proof states that $T^*$ must satisfy the identity

$$\bar{t}_{-i}(\theta_i, \delta_i) = \frac{\partial p_i(\delta_i)}{\partial \delta_i} - \bar{\pi}_{-i}(\theta_i, \delta_i) + \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds, \tag{A.38}$$

where $p_i : \Delta_i \to \mathbb{R}$ is some differentiable function. Conditions (A.11), (A.17), and (A.38) imply that

$$\begin{aligned} p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds &= U_i(\theta_i, \delta_i) \\ &= \bar{\pi}_i(\theta_i, \delta_i) + \bar{t}_i(\theta_i, \delta_i) + \delta_i \bar{\pi}_{-i}(\theta_i, \delta_i) + \delta_i \bar{t}_{-i}(\theta_i, \delta_i) \\ &= \bar{\pi}_i(\theta_i, \delta_i) + \bar{t}_i(\theta_i, \delta_i) + \delta_i \frac{\partial p_i(\delta_i)}{\partial \delta_i} + \delta_i \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds. \end{aligned} \tag{A.39}$$

Hence, $T^*$ must also satisfy the identity

$$\begin{aligned} \bar{t}_i(\theta_i, \delta_i) &= p_i(\delta_i) - \delta_i \frac{\partial p_i(\delta_i)}{\partial \delta_i} - \bar{\pi}_i(\theta_i, \delta_i) \\ &\quad + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds - \delta_i \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds. \end{aligned} \tag{A.40}$$

From identities (A.38) and (A.40), the transfer scheme $T^*$ can be "guessed". ■

## A.5. Proof of Theorem I.3 Continued

Consider the functions $S_i$ defined by (I.28) and the transfer scheme $T^*$ defined by (I.29) and (I.30). Notice first that, for all $(\hat{\theta}_1, \theta_2) \in \Theta$ and all $\delta \in \Delta$,

$$\mathbb{E}_{\theta_2}\big[t_1^*(\hat{\theta}_1, \theta_2, \delta) + \delta_1 t_2^*(\hat{\theta}_1, \theta_2, \delta)\big] = \mathbb{E}_{\theta_2}\big[S_1(\hat{\theta}_1, \theta_2, \delta)\big] - \mathbb{E}_{\theta_1,\theta_2}\left[S_1(\theta_1, \theta_2, \delta)\right], \tag{A.41}$$

$$\mathbb{E}_{\theta_2}\big[S_1(\hat{\theta}_1, \theta_2, \delta)\big] = \int_{\theta_1^{\min}}^{\hat{\theta}_1} \mathbb{E}_{\theta_2}\big[v_1(k^*(s, \theta_2, \delta))\big]\, ds - \mathbb{E}_{\theta_2}\big[\pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \hat{\theta}_1)\big] - \delta_1 \cdot \mathbb{E}_{\theta_2}\big[\pi_2(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_2\big], \tag{A.42}$$

where Fubini's theorem has been used to obtain equation (A.42). Under the assumption that agent 2 reveals her payoff type truthfully, agent 1 chooses $\hat{\theta}_1$ so as to maximize her interim expected utility. By making use of equations (A.41) and (A.42),

$$\begin{aligned}
&\mathbb{E}_{\theta_2}\left[u_1\big(k^*(\hat{\theta}_1, \theta_2, \delta), t_1^*(\hat{\theta}_1, \theta_2, \delta), t_2^*(\hat{\theta}_1, \theta_2, \delta), \theta_2 \,\big|\, \theta_1\big)\right] \\
=\; &\mathbb{E}_{\theta_2}\left[\big[\pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_1) + t_1^*(\hat{\theta}_1, \theta_2, \delta)\big] + \delta_1 \cdot \big[\pi_2(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_2) + t_2^*(\hat{\theta}_1, \theta_2, \delta)\big]\right] \\
=\; &\mathbb{E}_{\theta_2}\big[\pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_1)\big] + \delta_1 \cdot \mathbb{E}_{\theta_2}\big[\pi_2(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_2)\big] \\
&+ \mathbb{E}_{\theta_2}\big[S_1(\hat{\theta}_1, \theta_2, \delta)\big] - \mathbb{E}_{\theta_1,\theta_2}\left[S_1(\theta_1, \theta_2, \delta)\right] \\
=\; &\mathbb{E}_{\theta_2}\big[\pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_1)\big] - \mathbb{E}_{\theta_2}\big[\pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \hat{\theta}_1)\big] \\
&+ \int_{\theta_1^{\min}}^{\hat{\theta}_1} \mathbb{E}_{\theta_2}\big[v_1(k^*(s, \theta_2, \delta))\big]\, ds - \mathbb{E}_{\theta_1,\theta_2}\left[S_1(\theta_1, \theta_2, \delta)\right].
\end{aligned}$$

Hence, by making use of the Leibniz integral rule,

$$\begin{aligned}
&\frac{\partial}{\partial \hat{\theta}_1} \mathbb{E}_{\theta_2}\left[u_1\big(k^*(\hat{\theta}_1, \theta_2, \delta), t_1^*(\hat{\theta}_1, \theta_2, \delta), t_2^*(\hat{\theta}_1, \theta_2, \delta), \theta_2 \,\big|\, \theta_1\big)\right] \\
=\; &\mathbb{E}_{\theta_2}\left[\frac{\partial}{\partial \hat{\theta}_1} \pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \theta_1)\right] + \mathbb{E}_{\theta_2}\big[v_1(k^*(\hat{\theta}_1, \theta_2, \delta))\big] - \mathbb{E}_{\theta_2}\left[\frac{\partial}{\partial \hat{\theta}_1} \pi_1(k^*(\hat{\theta}_1, \theta_2, \delta) \,|\, \hat{\theta}_1)\right] \\
=\; &(\theta_1 - \hat{\theta}_1) \cdot \frac{\partial}{\partial \hat{\theta}_1} \mathbb{E}_{\theta_2}\left[v_1(k^*(\hat{\theta}_1, \theta_2, \delta))\right].
\end{aligned}$$

By assumption, the expected value in the last line is non-negative. Hence, truth-telling, $\hat{\theta}_1 = \theta_1$, maximizes agent 1's interim expected utility. By symmetry, $\hat{\theta}_2 = \theta_2$. As the above arguments hold for any set of (non-degenerate) type distributions, $T^*$ *strongly* Bayesian implements $k^*$. ■

## A.6. Derivation of the transfer scheme $T^*$ in the proof of Theorem I.3

Suppose externality types are common knowledge, and assume that the sensitive allocation rule $k^* : \Theta \times \Delta \to K$ is strongly Bayesian implemented by the ex post budget-balanced transfer scheme $T = (t_1, t_2) : \Theta \times \Delta \to \mathbb{R}^2$. Define

$$\begin{aligned}
\bar{v}_i(\hat{\theta}_i, \delta) &= \mathbb{E}_{\theta_{-i}}\big[v_i(k^*(\hat{\theta}_i, \theta_{-i}, \delta))\big],\\
\bar{h}_i(\hat{\theta}_i, \delta) &= \mathbb{E}_{\theta_{-i}}\big[h_i(k^*(\hat{\theta}_i, \theta_{-i}, \delta)\big],\\
\bar{\pi}_{-i}(\hat{\theta}_i, \delta) &= \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^*(\hat{\theta}_i, \theta_{-i}, \delta) \,|\, \theta_{-i})\big],\\
\bar{t}_i(\hat{\theta}_i, \delta) &= \mathbb{E}_{\theta_{-i}}\big[t_i(\hat{\theta}_i, \theta_{-i}, \delta)\big],\\
\bar{t}_{-i}(\hat{\theta}_i, \delta) &= \mathbb{E}_{\theta_{-i}}\big[t_{-i}(\hat{\theta}_i, \theta_{-i}, \delta)\big],
\end{aligned}$$

where $\pi_i(k \,|\, \theta_i) = \theta_i v_i(k) + h_i(k)$. For $i \in \{1, 2\}$, denote by $U_i(\hat{\theta}_i \,|\, \theta_i, \delta)$ agent $i$'s interim expected utility from reporting $\hat{\theta}_i$ if her true payoff type is $\theta_i$ and if agent $-i$ reports her payoff type truthfully:

$$U_i(\hat{\theta}_i \,|\, \theta_i, \delta) = \theta_i \bar{v}_i(\hat{\theta}_i, \delta) + \bar{h}_i(\hat{\theta}_i, \delta) + \bar{t}_i(\hat{\theta}_i, \delta) + \delta_i \bar{\pi}_{-i}(\hat{\theta}_i, \delta) + \delta_i \bar{t}_{-i}(\hat{\theta}_i, \delta) \tag{A.43}$$

By the same reasoning that has led to equation (A.17) in the proof Lemma I.1, the following must hold for all $i$ and all $(\theta_i, \delta)$:

$$U_i(\theta_i \,|\, \theta_i, \delta) = p_i(\delta) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta)\, ds \tag{A.44}$$

for some function $p_i : \Delta \to \mathbb{R}$. For ease of notation, write $t_i = t_i(\theta, \delta)$ and $\pi_i = \pi_i(k^*(\theta, \delta) \,|\, \theta_i)$. Then, by (A.44), the transfer scheme $T$ must satisfy the following identities:

$$\mathbb{E}_{\theta_2}\big[t_1\big] + \delta_1 \mathbb{E}_{\theta_2}\big[t_2\big] = p_1(\delta) + \int_{\theta_1^{\min}}^{\theta_1} \bar{v}_1(s,\delta)\, ds - \mathbb{E}_{\theta_2}\big[\pi_1\big] - \delta_1 \mathbb{E}_{\theta_2}\big[\pi_2\big], \tag{A.45}$$

$$\mathbb{E}_{\theta_1}\big[t_2\big] + \delta_2 \mathbb{E}_{\theta_1}\big[t_1\big] = p_2(\delta) + \int_{\theta_2^{\min}}^{\theta_2} \bar{v}_2(s,\delta)\, ds - \mathbb{E}_{\theta_1}\big[\pi_2\big] - \delta_2 \mathbb{E}_{\theta_1}\big[\pi_1\big]. \tag{A.46}$$

Due to budget balance, identities (A.45) and (A.46) imply that the interim expected transfers must satisfy

$$(1-\delta_1)\mathbb{E}_{\theta_2}\big[t_1\big] = p_1(\delta) + \int_{\theta_1^{\min}}^{\theta_1} \bar{v}_1(s,\delta)\, ds - \mathbb{E}_{\theta_2}\big[\pi_1\big] - \delta_1 \mathbb{E}_{\theta_2}\big[\pi_2\big],$$

$$-(1-\delta_2)\mathbb{E}_{\theta_1}\big[t_1\big] = p_2(\delta) + \int_{\theta_2^{\min}}^{\theta_2} \bar{v}_2(s,\delta)\, ds - \mathbb{E}_{\theta_1}\big[\pi_2\big] - \delta_2 \mathbb{E}_{\theta_1}\big[\pi_1\big],$$

and

$$-(1-\delta_1)\mathbb{E}_{\theta_2}\big[t_2\big] = p_1(\delta) + \int_{\theta_1^{\min}}^{\theta_1} \bar{v}_1(s,\delta)\, ds - \mathbb{E}_{\theta_2}\big[\pi_1\big] - \delta_1 \mathbb{E}_{\theta_2}\big[\pi_2\big],$$

$$(1-\delta_2)\mathbb{E}_{\theta_1}\big[t_2\big] = p_2(\delta) + \int_{\theta_2^{\min}}^{\theta_2} \bar{v}_2(s,\delta)\, ds - \mathbb{E}_{\theta_1}\big[\pi_2\big] - \delta_2 \mathbb{E}_{\theta_1}\big[\pi_1\big],$$

whereas ex post budget balance requires that also $t_1 + t_2 = 0$. From these conditions, the transfer scheme $T^*$ can be "guessed". ■

## A.7. Proof of Proposition I.5 Continued

Implicit differentiation of (I.33) yields: $\frac{\partial k^*}{\partial \delta_1} = \frac{\theta_2 v(1-k^*)}{\partial F/\partial k^*}$; $\frac{\partial k^*}{\partial \delta_2} = \frac{-\theta_1 v(k^*)}{\partial F/\partial k^*}$; $\frac{\partial k^*}{\partial \theta_1} = \frac{-v(1-k^*)-\delta_2 v(k^*)}{\partial F/\partial k^*}$; and $\frac{\partial k^*}{\partial \theta_2} = \frac{\delta_1 v(1-k^*)+v(k^*)}{\partial F/\partial k^*}$.

Notice that $\frac{\partial F}{\partial k^*} = -(\theta_1 - \delta_1\theta_2)v'(1-k^*) - (\theta_2 - \delta_2\theta_1)v'(k^*) < 0$, since $(\theta_i - \delta_i\theta_{-i}) > 0$ as $\delta_i^{\max} < \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}$. By substituting for (I.33), $v(1-k^*) + \delta_2 v(k^*) = \frac{1-\delta_1\delta_2}{\theta_1-\delta_1\theta_2}\theta_2 v(k^*) > 0$ and $\delta_1 v(1-k^*) + v(k^*) = \frac{1-\delta_1\delta_2}{\theta_1-\delta_1\theta_2}\theta_1 v(k^*) > 0$. Hence, $\frac{\partial k^*}{\partial \delta_1} < 0 < \frac{\partial k^*}{\partial \delta_2}$ and $\frac{\partial k^*}{\partial \theta_2} < 0 < \frac{\partial k^*}{\partial \theta_1}$.

Implicit differentiation of (I.34) yields: $\frac{\partial k^*}{\partial \delta_1} = \frac{\theta_2^2 v(1-k^*)}{\partial G/\partial k^*}$; $\frac{\partial k^*}{\partial \delta_2} = \frac{-\theta_1^2 v(k^*)}{\partial G/\partial k^*}$; $\frac{\partial k^*}{\partial \theta_1} = \frac{(\theta_2-2\delta_2\theta_1)v(k^*)-\theta_2 v(1-k^*)}{\partial G/\partial k^*}$; and $\frac{\partial k^*}{\partial \theta_2} = \frac{\theta_1 v(k^*)-(\theta_1-2\delta_1\theta_2)v(1-k^*)}{\partial G/\partial k^*}$.

Notice that $\frac{\partial G}{\partial k^*} = -\theta_2(\theta_1 - \delta_1\theta_2)v'(1-k^*) - \theta_1(\theta_2 - \delta_2\theta_1)v'(k^*) < 0$, since $(\theta_i - \delta_i\theta_{-i}) > 0$ as $\delta_i^{\max} < \frac{\theta_i^{\min}}{\theta_{-i}^{\max}}$. By substituting for (I.34),

$$\begin{aligned}
& (\theta_2 - 2\delta_2\theta_1)v(k^*) - \theta_2 v(1-k^*) \\
= \quad & -(\delta_1\theta_2^2 - 2\delta_1\delta_2\theta_1\theta_2 + \delta_2\theta_1^2)\frac{v(k^*)}{\theta_1 - 2\delta_1\theta_2} \\
= \quad & -\left[(\delta_1\theta_2 - \delta_2\theta_1)^2 + \delta_1(1-\delta_1)\theta_2^2 + \delta_2(1-\delta_2)\theta_1^2\right]\frac{v(k^*)}{\theta_1 - 2\delta_1\theta_2} \\
= \quad & -\left[\theta_1 v(k^*) - (\theta_1 - 2\delta_1\theta_2)v(1-k^*)\right].
\end{aligned} \tag{A.47}$$

Hence, $\operatorname{sgn}(\frac{\partial k^*}{\partial \theta_1}\frac{\partial k^*}{\partial \theta_2}) = -1 = \operatorname{sgn}(\frac{\partial k^*}{\partial \delta_1}\frac{\partial k^*}{\partial \delta_2})$. When assuming $\Delta_i \subset [0, \frac{\theta_i^{\min}}{2\theta_{-i}^{\max}}]$, the term in the third line of (A.47) is negative. In this case, $\frac{\partial k^*}{\partial \theta_2} < 0 < \frac{\partial k^*}{\partial \theta_1}$. ∎

# Appendix B.

# Appendix to Chapter II

For a proof of Proposition II.3, it is convenient to reformulate Proposition II.2. For this purpose, let $x_{i,(ij)}$ denote $i$'s effort in assignment $[(ij)(kl)]$, as given by (II.6), and let $x_{i,(abcd)}$ denote $i$'s effort in assignment $[(abcd)]$, as given by (II.13).

**Lemma B.1** *Let $w$ be a fixed piece rate for both teams. Suppose there is a combination of agents, $\{i,j,k,l\} = \{a,b,c,d\}$, such that $m_{ij} + m_{kl} \geq M_i + M_k, M_j + M_l$. And suppose that one of the following conditions is satisfied.*

*(i) Marginal costs of effort are concave, and*

$$\begin{aligned}\max\{m_{ij}, m_{kl}\} &> \max\{\max\{M_i, 0\}, \max\{M_k, 0\}\},\\ &\quad \max\{\max\{M_j, 0\}, \max\{M_l, 0\}\}. \qquad \text{(B.1)}\end{aligned}$$

*(ii) Marginal costs of effort are convex, and*

$$\begin{aligned}\min\{m_{ij}, m_{kl}\} &> \min\{\max\{M_i, 0\}, \max\{M_k, 0\}\},\\ &\quad \min\{\max\{M_j, 0\}, \max\{M_l, 0\}\}. \qquad \text{(B.2)}\end{aligned}$$

*Then* $\Pi_{[(ij)(kl)]} > \Pi_{[(abcd)]}$.

**Proof.** The assumptions imply that $x_{i,(ij)} + x_{k,(kl)} > x_{i,(abcd)} + x_{k,(abcd)}$ and $x_{j,(ij)} + x_{l,(kl)} > x_{j,(abcd)} + x_{l,(abcd)}$. Hence, $\Pi_{[(ij)(kl)]} > \Pi_{[(abcd)]}$. ■

## B.1. Proof of Proposition II.3

Suppose throughout $M_a \geq M_b \geq M_c \geq M_d$. I distinguish between the following cases.

**(A)** $0 \geq M_a \geq M_b \geq M_c \geq M_d$.

**(B)** $M_a > 0 \geq M_b \geq M_c \geq M_d$, or $M_a \geq M_b > 0 \geq M_c \geq M_d$.

**(C)** $M_a \geq M_b \geq M_c > 0 \geq M_d$, and $C_{xxx} < 0$ .

**(D)** $M_a \geq M_b \geq M_c \geq M_d > 0$, and $C_{xxx} < 0$ .

**(E)** $M_a \geq M_b \geq M_c \geq M_d > 0$, and $C_{xxx} > 0$.

The case $M_a \geq M_b \geq M_c > 0 \geq M_d$, with $C_{xxx} > 0$, has been discussed in Section II.4.

**Ad (A):** By (II.7) and (II.14), $\Pi_{[(ij)(kl)]} \geq 0 = \Pi_{[(abcd)]}$.

**Ad (B):** By (II.13), $c$ and $d$ exert zero effort in $[(abcd)]$. Let $j \in \{b, c, d\}$ such that $m_{aj} = \max\{m_{ab}, m_{ac}, m_{ad}\}$, and let $\{k, l\} = \{b, c, d\} \setminus \{j\}$. Then, either $m_{aj} > M_a$ or $m_{ab} = m_{ac} = m_{ad} = 1$. If $m_{aj} > M_a \geq M_b$, Lemma B.1 yields

$$\begin{aligned} \Pi_{[(aj)(kl)]} &\geq (1-w)\left[C_x^{-1}(m_{aj}w) + C_x^{-1}(m_{aj}w)\right] \\ &> (1-w)\left[C_x^{-1}(w\max\{M_a, 0\}) + C_x^{-1}(w\max\{M_b, 0\})\right] \\ &= \Pi_{[(abcd)]}. \end{aligned} \tag{B.3}$$

If $m_{ab} = m_{ac} = m_{ad} = 1$, then $\delta_{ad} = 1$. In this case, $4M_d \geq 2 + \delta_{bd} + \delta_{cd} \geq 0$. By assumption, $0 \geq M_c \geq M_d$, thus, $M_c = M_d = 0$. Together, $\delta_{ab} = \delta_{ac} = \delta_{ad} = 1$ and $\delta_{bc} = \delta_{bd} = \delta_{cd} = -1$. But then, also $M_b = 0$. Thus, for all $\{j, k, l\} = \{b, c, d\}$, $\Pi_{[(aj)(kl)]} = 2(1-w)C_x^{-1}(1 \cdot w) > (1-w)C_x^{-1}(1 \cdot w) = \Pi_{[(abcd)]}$.

**Ad (C):** The conditions of this case imply that $1 > M_a$: If $M_a = 1$, then $\delta_{ad} = 1$, and $M_d \geq 0$, a contradiction. Again, $d$ exerts zero effort. Obviously, $m_{aj} < M_a$ for at most one $j \neq a$. I distinguish between (I) $m_{ab}, m_{ac} \geq M_a$, (II) $m_{ab}, m_{ad} \geq M_a$, and (III) $m_{ac}, m_{ad} \geq M_a$.

Consider (I). Assume that $m_{ab} + m_{cd} \leq M_b + M_c$ and $m_{ac} + m_{bd} \leq M_b + M_c$. Summing up both inequalities yields $\delta_{bc} \geq 1$. If $\delta_{bc} < 1$, then this implies that $m_{ab} + m_{cd} > M_b + M_c$ or $m_{ac} + m_{bd} > M_b + M_c$. By assumption, $m_{ab}, m_{ac} \geq \max\{M_a, M_b, M_c, M_d\}$. Obviously, $m_{ab} + m_{cd}, m_{ac} + m_{bd} \geq M_a + 0$. Lemma B.1(i) implies that $\Pi_{[(ab)(cd)]} > \Pi_{[(abcd)]}$ or $\Pi_{[(ac)(bd)]} > \Pi_{[(abcd)]}$. If $\delta_{bc} = 1$ instead, then $m_{bc} = 1 > M_a = \max\{M_a, M_b, M_c, M_d\}$ . Obviously, $m_{bc} + m_{ad} = 1 + m_{ad} > M_c + 0$. From $M_a \geq M_b$ and $\delta_{bc} = 1$ follows $\delta_{ad} \geq \delta_{bd}$. With this, $m_{bc} + m_{ad} \geq M_a + M_b$. By Lemma B.1(i), $\Pi_{[(ad)(bc)]} > \Pi_{[(abcd)]}$.

Consider (II). Assume that $m_{ab} + m_{cd} \leq M_a + M_c$ and $m_{ad} + m_{bc} \leq M_a + M_c$. Summing up both inequalities yields $\delta_{ac} \geq 1$. If $\delta_{ac} = 1$, case (I) applies. If $\delta_{ac} < 1$, then $m_{ab} + m_{cd} > M_a + M_c$ or $m_{ad} + m_{bc} > M_a + M_c$. By assumption, $m_{ab}, m_{ad} \geq \max\{M_a, M_b, M_c, M_d\}$ . Obviously, $m_{ab} + m_{cd}, m_{ad} + m_{bc} \geq M_b + 0$. By Lemma B.1(i), $\Pi_{[(ab)(cd)]} > \Pi_{[(abcd)]}$ or $\Pi_{[(ad)(bc)]} > \Pi_{[(abcd)]}$.

Consider (III). Assume that $m_{ac} + m_{bd} \leq M_a + M_b$ and $m_{ad} + m_{bc} \leq M_a + M_b$. Summing up both inequalities yields $\delta_{ab} \geq 1$. If $\delta_{ab} = 1$, case (I) applies. If $\delta_{ab} < 1$, then $m_{ac} + m_{bd} > M_a + M_b$ or $m_{ad} + m_{bc} > M_a + M_b$. By assumption, $m_{ac}, m_{ad} \geq \max\{M_a, M_b, M_c, M_d\}$ . Obviously, $m_{ac} + m_{bd}, m_{ad} + m_{bc} \geq M_c + 0$. By Lemma B.1(i), $\Pi_{[(ac)(bd)]} > \Pi_{[(abcd)]}$ or $\Pi_{[(ad)(bc)]} > \Pi_{[(abcd)]}$.

**Ad (D):** Assume, without loss of generality, that $\delta_{ij} + \delta_{kl} \geq \delta_{ik} + \delta_{jl} \geq \delta_{il} + \delta_{jk}$ and $\delta_{ij} \geq \delta_{kl}$ for some combination of agents, $\{i, j, k, l\} = \{a, b, c, d\}$.

We have $m_{ij} + m_{kl} \geq M_i + M_k$, since $2 + \delta_{ij} + \delta_{kl} \geq 2\delta_{ik} + \delta_{il} + \delta_{jk}$ by $\delta_{ij} + \delta_{kl} \geq \delta_{il} + \delta_{jk}$ and $1 \geq \delta_{ik}$. Similarly, $m_{ij} + m_{kl} \geq M_j + M_l$. If, in addition, $\max\{m_{ij}, m_{kl}\} = m_{ij} \geq \max\{M_i, M_j, M_k, M_l\}$, then Lemma B.1(i) yields $\Pi_{[(ij)(kl)]} \geq \Pi_{[(abcd)]}$, where $\Pi_{[(ij)(kl)]} = \Pi_{[(abcd)]}$ holds if and only if $\delta_{ij} = 1$ for all $i, j \in \{a, b, c, d\}$. I show in the following that $\Pi_{[(ik)(jl)]} > \Pi_{[(abcd)]}$ if $m_{ij} < \max\{M_i, M_j, M_k, M_l\}$.

If $m_{ij} < \max\{M_i, M_j, M_k, M_l\}$, then at least one of the following inequalities does not hold: (I) $m_{ij} \geq M_i$, (II) $m_{ij} \geq M_j$, (III) $m_{ij} \geq M_k$, (IV) $m_{ij} \geq M_l$. On the other hand, at most one of these inequalities does not hold. To see this, assume that, for instance, (III) and (IV) are not satisfied. Then, adding up yields $m_{ij} + m_{ij} < M_k + M_l$, such that $2 + 4\delta_{ij} < \delta_{ik} + \delta_{il} + \delta_{jk} + \delta_{jl} + 2\delta_{kl}$. Therefore, $4\delta_{ij} < \delta_{ik} + \delta_{il} + \delta_{jk} + \delta_{jl}$, which contradicts $\delta_{ij} + \delta_{kl} \geq \delta_{ik} + \delta_{jl} \geq \delta_{il} + \delta_{jk}$, and $\delta_{ij} \geq \delta_{kl}$. The argument is similar for any other two inequalities among (I)-(IV). Furthermore, $m_{ik}+m_{jl} \geq M_i+M_j$ since $2+\delta_{ik}+\delta_{jl} \geq 2\delta_{ij}+\delta_{il}+\delta_{jk}$ by $\delta_{ik}+\delta_{jl} \geq \delta_{il}+\delta_{jk}$ and $1 \geq \delta_{ij}$. Similarly, $m_{ik}+m_{jl} \geq M_k+M_l$. I show that $\max\left\{m_{ik}, m_{jl}\right\} > \max\left\{M_i, M_j, M_k, M_l\right\}$ in all four cases in which exactly one inequality among (I)-(IV) does not hold. Lemma B.1(i) then implies that $\Pi_{[(ik)(jl)]} > \Pi_{[(abcd)]}$.

Assume (I) does not hold, $m_{ij} < M_i$. Then, $1 + \delta_{ij} < \delta_{ik} + \delta_{il}$. Therefore, $\delta_{ij} < \delta_{ik}, \delta_{il}$, since $\delta_{ik}, \delta_{il} \leq 1$. By $\delta_{ij} + \delta_{kl} \geq \delta_{ik} + \delta_{jl}$ and $\delta_{ij} < \delta_{ik}$ we have $\delta_{kl} > \delta_{jl}$. By $\delta_{ij} + \delta_{kl} \geq \delta_{il} + \delta_{jk}$ and $\delta_{ij} < \delta_{il}$, we have $\delta_{kl} > \delta_{jk}$. Together, $\delta_{ik}, \delta_{il} > \delta_{ij} \geq \delta_{kl} > \delta_{jk}, \delta_{jl}$. With this, it is easy to see that $\max\left\{m_{ik}, m_{jl}\right\} = m_{ik} > \max\left\{M_i, M_j, M_k, M_l\right\}$.

Assume (II) does not hold, $m_{ij} < M_j$. Then $\delta_{jk}, \delta_{jl} > \delta_{ij} \geq \delta_{kl} > \delta_{ik}, \delta_{il}$ and $\max\left\{m_{ik}, m_{jl}\right\} = m_{jl} > \max\left\{M_i, M_j, M_k, M_l\right\}$.

Assume (III) does not hold, $m_{ij} < M_k$. Then, $1 + 2\delta_{ij} < \delta_{ik} + \delta_{jk} + \delta_{kl}$. Thus $1 + \delta_{ij} < \delta_{ik} + \delta_{jk} + \delta_{kl} - \delta_{ij} \leq \delta_{ik} + \delta_{jk}$. Therefore, $\delta_{ij} < \delta_{ik}, \delta_{jk}$ since $\delta_{ik}, \delta_{jk} \leq 1$. Similarly as above, $\delta_{ik}, \delta_{jk} > \delta_{ij} \geq \delta_{kl} > \delta_{il}, \delta_{jl}$. With this, $\max\left\{m_{ik}, m_{jl}\right\} = m_{ik} > \max\left\{M_i, M_j, M_k, M_l\right\}$ .

Assume (IV) does not hold, $m_{ij} < M_l$. Then $\delta_{il}, \delta_{jl} > \delta_{ij} \geq \delta_{kl} > \delta_{ik}, \delta_{jk}$. With this, $\max\left\{m_{ik}, m_{jl}\right\} = m_{jl} > \max\left\{M_i, M_j, M_k, M_l\right\}$.

In all four cases, Lemma B.1(i) yields $\Pi_{[(ik)(jl)]} > \Pi_{[(abcd)]}$.

**Ad (E):** Assume, without loss of generality, that $\delta_{ij} + \delta_{kl} \geq \delta_{ik} + \delta_{jl} \geq \delta_{il} + \delta_{jk}$ and $\delta_{ij} \geq \delta_{kl}$ for some combination of agents, $\{i, j, k, l\} = \{a, b, c, d\}$. This yields $2\left(1 + \delta_{ij}\right) \geq 2\left(\delta_{ij} + \delta_{kl}\right) \geq \left(\delta_{ik} + \delta_{jl}\right) + \left(\delta_{il} + \delta_{jk}\right)$. Thus, $m_{ij} + m_{ij} \geq M_i + M_j$ and $m_{kl} + m_{kl} \geq M_k + M_l$. Therefore, $m_{ij} \geq \min\left\{M_i, M_j\right\}$ and $m_{kl} \geq \min\left\{M_k, M_l\right\}$. By Lemma B.1(ii), $\Pi_{[(ij)(kl)]} \geq \Pi_{[(abcd)]}$, where $\Pi_{[(ij)(kl)]} = \Pi_{[(abcd)]}$ if and only if $\delta_{ij} = 1$ for all $i, j \in \{a, b, c, d\}$. ■

# Appendix C.

# Appendix to Chapter III

## C.1. Proof of Lemma III.1

**Ad (i):** Define the function $\hat{f}(y) = y + f(y)$, and let $\hat{y}$ such that $\hat{f}(\hat{y}) = \frac{\phi}{1-\phi} I_P$. Since $y + f(y) > I_P$ by assumption, and $\frac{\phi}{1-\phi} \leq 1$, $\hat{y}$ exists; as $\hat{f}$ increases in $y$, $\hat{y}$ is unique. Suppose the child could legally earn $y = \hat{y}$ but chooses to engage entirely in illegal activities, $x = \hat{y}$. If he remains undetected, (III.4) implies that his parent is then indifferent on the margin between providing him with a transfer or not. For any ex post disposable income $I_C^i < \hat{f}(\hat{y})$, with $i \in \{d, u\}$, the parent therefore strictly prefers to support her child. This establishes regime $R1$ as the relevant one for legal income prospects $y \leq \hat{y}$, since then $I_C^i \leq \hat{f}(y) \leq \hat{f}(\hat{y})$ for any level of illegal activity.

**Ad (iii):** Notice that $\hat{y} < \frac{\phi}{1-\phi} I_P$. Suppose the child fully abides by the laws and earns $y = \frac{\phi}{1-\phi} I_P$, such that $I_C^u = y = I_C^d$. By (III.4), his parent is then indifferent on the margin between providing him with a transfer or not. Since illegal activity raises the child's disposable income when undetected, but decreases it when detected, any illegal effort $x > 0$ leads the parent to provide her child with a transfer if and only if he is detected. This establishes regime $R2$ as the relevant one if the child's legal income prospects amount to $y = \frac{\phi}{1-\phi} I_P$, regardless of the illegal effort he chooses.

**Ad (v):** Suppose the child could earn a maximum legal income of $y = \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P$, but engages fully in illegal activities and is detected, such that $I_C^d = y - \pi g(y) = (1-\pi)y$. By (III.4), his parent is then indifferent on the margin between providing him with a transfer or not. This establishes regime $R3$ as the relevant one if the child's legal income prospects amount to $y \geq \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P$, since then $I_C^u \geq I_C^d > (1-\pi)y$ for any level of illegal activity.

**Ad (ii) and (iv):** Consider legal income prospects $y \in [\hat{y}, \frac{\phi}{1-\phi}I_P]$. By the arguments in part (i), regime $R2$ potentially applies as soon as $y$ marginally increases above $\hat{y}$, with a marginal transfer only in case the child engages nearly entirely in illegal activities and is detected. As the child's optimum crime level at an income of $\hat{y}$ constitutes an interior solution under regime $R1$, there exists some $\epsilon > 0$ such that, for legal income prospects $y \in [\hat{y}, \hat{y} + \epsilon]$, the child strictly prefers regime $R1$ over $R2$ and chooses $x$ accordingly.

By applying a similar argument to the critical income prospects of parts (iii) and (v), there exists some $\epsilon > 0$ such that the following holds: For income prospects $y \in [\frac{\phi}{1-\phi}I_P - \epsilon, \frac{\phi}{1-\phi}I_P]$, the child strictly prefers regime $R2$ over $R1$; for income prospects $y \in [\frac{\phi}{1-\phi}I_P, \frac{\phi}{1-\phi}I_P + \epsilon]$, the child strictly prefers regime $R2$ over $R3$; and for income prospects $y \in [\frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P - \epsilon, \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P]$, the child strictly prefers regime $R3$ over $R2$.

Now consider the child's indirect expected utility function

$$V(s_C, y, p, \pi | Rj) = \max_x E[U_C | Rj], \tag{C.1}$$

where $Rj$ denotes the relevant regime. By taking derivatives of the child's indirect utility (III.8) with respect to his maximum legal income $y$, while substituting for his first-order conditions (III.11)-(III.13), we obtain:[1]

$$V_y(\cdot|R1) = \phi(1-p)u'(c^u_{C|R1}) + \phi p u'(c^d_{C|R1}), \tag{C.2}$$

$$V_y(\cdot|R2) = (1-p)u'(c^u_{C|R2}) + \phi p u'(c^d_{C|R2}), \tag{C.3}$$

$$V_y(\cdot|R3) = (1-p)u'(c^u_{C|R3}) + p u'(c^d_{C|R3}). \tag{C.4}$$

[1] We denote the partial derivatives $\frac{\partial}{\partial\lambda}V(\cdot|Rj)$, for $\lambda \in \{s_C, y, p, \pi\}$, by $V_\lambda(\cdot|Rj)$. We denote by $c^i_{C|Rj}$ the child's equilibrium consumption under regime $Rj$ in state $i \in \{u, d\}$.

It is straight forward to show that $V_y(\cdot|R1) - V_y(\cdot|R2) < 0$, and $V_y(\cdot|R2) - V_y(\cdot|R3) < 0$. Hence, there exist unique income prospects $\{y_1^*, y_2^*\}$ as claimed. ■

## C.2. Proof of Proposition III.1

We prove the Proposition in a comprehensive way. The four critical levels of the child's legal income prospects, $\{\tilde{y}_{11}, \tilde{y}_{12}, \tilde{y}_{23}, \tilde{y}_{33}\}$, are defined as follows.

- $\tilde{y}_{11}$ is defined as the maximum legal income at which the child is indifferent between his optimum crime level $x_{R1}$ under regime $R1$ and his optimum crime level $x_{R2}$ under regime $R2$, provided the parent instills in him the equilibrium ethic associated with interior solutions under regime $R1$, as given by (III.16). Formally: With $y_1^* = y_1^*(s_C)$ as characterized by Lemma III.1, $\tilde{y}_{11} = y_1^*(\frac{\phi}{\alpha} s_P)$. Furthermore, for $y \leq \tilde{y}_{11}$, define

$$s_C^*(y) = \frac{\phi}{\alpha} s_P. \tag{C.5}$$

- $\tilde{y}_{12}$ is defined as the maximum legal income at which the child is indifferent between his optimum crime level $x_{R1}$ under regime $R1$ and his optimum crime level $x_{R2}$ under regime $R2$, provided the parent instills in him the equilibrium ethic associated with interior solutions under regime $R2$, as given by (III.17). Formally: With $y_1^* = y_1^*(s_C)$ as characterized by Lemma III.1, $\tilde{y}_{12} = y_1^*(s_C^*(\tilde{y}_{12}))$ for

$$s_C^*(\tilde{y}_{12}) = \frac{1}{\alpha}[s_P + (1 - \phi)u'(c_P^d)]. \tag{C.6}$$

- $\tilde{y}_{23}$ is defined as the maximum legal income at which the child is indifferent between his optimum crime level $x_{R2}$ under regime $R2$ and his optimum crime level $x_{R3}$ under regime $R3$, provided the parent instills in him the equilibrium ethic associated with interior solutions under regime $R2$, as given by (III.17). Formally: With $y_2^* = y_2^*(s_C)$ as characterized by Lemma III.1, $\tilde{y}_{23} = y_2^*(s_C^*(\tilde{y}_{23}))$ for

$$s_C^*(\tilde{y}_{23}) = \frac{1}{\alpha}[s_P + (1 - \phi)u'(c_P^d)]. \tag{C.7}$$

- $\tilde{y}_{33}$ is defined as the maximum legal income at which the child is indifferent between his optimum crime level $x_{R2}$ under regime $R2$ and his optimum crime level $x_{R3}$ under regime $R3$, provided the parent instills in him the equilibrium ethic associated with interior solutions under regime $R3$, as given by (III.18). Formally: With $y_2^* = y_2^*(s_C)$ as characterized by Lemma III.1, $\tilde{y}_{33} = y_2^*(\frac{1}{\alpha}s_P)$. Furthermore, for $y \geq \tilde{y}_{33}$, define

$$s_C^*(y) = \frac{1}{\alpha}s_P. \tag{C.8}$$

In what follows, we denote the parent's expected utility per regime by

$$\begin{aligned} E[U_P|R1] &= (1-p)u\left(c^u_{P|R1}\right) + pu\left(c^d_{P|R1}\right) - ps_P\pi g(x_{R1}) & \text{(C.9)} \\ &\quad +\alpha\left[V(\cdot|R1) + ps_C\pi g(x_{R1})\right], \\ E[U_P|R2] &= (1-p)u(I_P) + pu\left(c^d_{P|R2}\right) - ps_P\pi g(x_{R2}) & \text{(C.10)} \\ &\quad +\alpha\left[V(\cdot|R2) + ps_C\pi g(x_{R2})\right], \\ E[U_P|R3] &= u(I_P) - ps_P\pi g(x_{R3}) & \text{(C.11)} \\ &\quad +\alpha\left[V(\cdot|R3) + ps_C\pi g(x_{R3})\right], \end{aligned}$$

where $V(\cdot|Rj)$ denotes the child's regime-dependent indirect utility from (C.1).

We first prove the Proposition for legal income prospects $y \leq \frac{\phi}{1-\phi}I_P$. We know from (III.16) and Lemma III.1(i) that, for $y \leq \hat{y}$, equilibrium ethic is given by $s_C^*(y) = \frac{\phi}{\alpha}s_P$. Furthermore, we know from Lemma III.1 that, for $s_C^*(y) = \frac{\phi}{\alpha}s_P$, there exists a legal income level $y_1^*(\frac{\phi}{\alpha}s_P) \in (\hat{y}, \frac{\phi}{1-\phi}I_P)$ such that the child prefers $R1$ over $R2$ for all $y \leq y_1^*(\frac{\phi}{\alpha}s_P)$, while changing preferences at $y = y_1^*(\frac{\phi}{\alpha}s_P)$. By definition, $\tilde{y}_{11} = y_1^*(\frac{\phi}{\alpha}s_P)$. We can contrast the child's preferences with the parent's by considering the difference

$$\begin{aligned} E[U_P|R1] - E[U_P|R2] &= \left(\frac{1-\phi}{\phi}\right)^{1-\rho}\left[V(\cdot|R1) - V(\cdot|R2)\right] & \text{(C.12)} \\ &\quad + p\pi\left[g(x_{R2}) - g(x_{R1})\right]\left(s_P - \frac{\alpha}{\phi}s_C\right) \\ &\quad + (1-p)\left(\left(\frac{1-\phi}{\phi}\right)^{1-\rho}u(c^u_{C|R2}) - u(I_P)\right), \end{aligned}$$

where we have substituted for the identity $u(c^i_{P|Rj}) = u\left(\left(\frac{1-\phi}{\phi}\right)c^i_{C|Rj}\right) = \left(\frac{1-\phi}{\phi}\right)^{1-\rho} u(c^i_{C|Rj})$, which holds if transfers are operative. As $I_P < \left(\frac{1-\phi}{\phi}\right)c^u_{C|R2}$ according to (III.4), the term in the last line of (C.12) is positive. Hence, for $s_C = s^*_C = \frac{\phi}{\alpha}s_P$,

$$E\left[U_P|R1\right] - E\left[U_P|R2\right] \quad > \quad \left(\frac{1-\phi}{\phi}\right)^{1-\rho}\left[V\left(\cdot|R1\right) - V\left(\cdot|R2\right)\right]. \tag{C.13}$$

That is, as long as the child *weakly* prefers $R1$ over $R2$, the parent *strictly* prefers $R1$ over $R2$. Specifically, there exists some $\epsilon_1 > 0$ such that the parent strictly prefers $R1$ over $R2$ for all legal income prospects $y \le y^*_1(\frac{\phi}{\alpha}s_P) + \epsilon_1$.

In order to make $R1$ incentive compatible for her child, she instills a stronger ethic in him, $s^*_C > \frac{\phi}{\alpha}s_P$, so as to keep him just indifferent between regimes $R1$ and $R2$: On the one hand, she does not want to increase $s^*_C$ beyond indifference of her child, since, according to (III.16), her marginal expected utility with respect to $s_C$ decreases for $s_C > \frac{\phi}{\alpha}s_P$. On the other hand, she can indeed enforce indifference: By (III.8), $V_{s_C}(\cdot|Rj) = -px_{Rj} < 0$, whereas we know from (III.11) and (III.12) that, *ceteris paribus*, $x_{R1} < x_{R2}$; hence, $V_{s_C}(\cdot|R1) > V_{s_C}(\cdot|R2)$. In particular, for income prospects $y \in (y^*_1(\frac{\phi}{\alpha}s_P), y^*_1(\frac{\phi}{\alpha}s_P) + \epsilon_1)$,

$$\frac{ds^*_C}{dy} = -\frac{V_y(\cdot|R1) - V_y(\cdot|R2)}{V_{s_C}(\cdot|R1) - V_{s_C}(\cdot|R2)} > 0, \tag{C.14}$$

where we make use of the fact that regime $R2$ gets even more attractive to the child as his legal income prospects increase: $V_y(\cdot|R1) - V_y(\cdot|R2) < 0$. Consequently, $s^*_C(y)$ is determined through $V(\cdot|R1) = V(\cdot|R2)$ for increasing legal income prospects $y > y^*_1(\frac{\phi}{\alpha}s_P)$ *at least* as long as the parent does not herself prefer $R2$ over $R1$.

By a similar argument as in the proof of Lemma III.1(ii), there exists some $\epsilon_2 > 0$ such that, for the equilibrium ethic determined by (III.17), both parent and child prefer regime $R2$ over $R1$ if $y \in (\frac{\phi}{1-\phi}I_P - \epsilon_2, \frac{\phi}{1-\phi}I_P]$. Hence, there exists a maximum $\hat{\epsilon}_1 > 0$ such that the parent (weakly) prefers $R1$ over $R2$ for all income prospects $y \le y^*_1(\frac{\phi}{\alpha}s_P) + \hat{\epsilon}_1$, but prefers $R2$ over $R1$ if $y \in (y^*_1(\frac{\phi}{\alpha}s_P) + \hat{\epsilon}_1, \frac{\phi}{1-\phi}I_P]$. Similarly, there exists a maximum $\hat{\epsilon}_2 > 0$ such that, for the equilibrium ethic determined by (III.17), both parent and child (weakly) prefer regime $R2$ over $R1$ if $y \in [\frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2, \frac{\phi}{1-\phi}I_P]$. We show in the following that

$y_1^*(\frac{\phi}{\alpha}s_P) + \hat{\epsilon}_1 < \frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2$, and that, for income prospects $y \in (y_1^*(\frac{\phi}{\alpha}s_P) + \hat{\epsilon}_1, \frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2)$, equilibrium ethic is (still) determined through $V(\cdot|R1) = V(\cdot|R2)$.

For this purpose, consider the ethic $s_C^*$ corresponding to interior solutions under regime $R2$. Taking differences of the parent's expected utility for regimes $R1$ and $R2$, while substituting for $s_C^* = \frac{1}{\alpha}s_P + \frac{1}{\alpha}(1-\phi)u'(c_P^d)$, yields

$$\begin{aligned} E[U_P|R1] - E[U_P|R2] &= (1-p)u(c^u_{P|R1}) + pu(c^d_{P|R1}) \\ &\quad - (1-p)u(I_P) - pu(c^d_{P|R2}) \\ &\quad - pu'(c^d_{P|R2})(1-\phi)\pi[g(x_{R2}) - g(x_{R1})] \\ &\quad + \alpha[V(\cdot|R1) - V(\cdot|R2)]. \end{aligned} \tag{C.15}$$

Concavity of utility from consumption implies that

$$\begin{aligned} u\left(c^d_{P|R1}\right) - u\left(c^d_{P|R2}\right) &< u'(c^d_{P|R2})(c^d_{P|R1} - c^d_{P|R2}) \\ &= u'(c^d_{P|R2})(1-\phi)\pi[g(x_{R2}) - g(x_{R1})]. \end{aligned} \tag{C.16}$$

Substituting (C.16) into (C.15) yields

$$\begin{aligned} E[U_P|R1] - E[U_P|R2] &< \alpha[V(\cdot|R1) - V(\cdot|R2)] \\ &\quad - (1-p)[u(I_P) - u(c^u_{P|R1})] \\ &< \alpha[V(\cdot|R1) - V(\cdot|R2)], \end{aligned} \tag{C.17}$$

where the second inequality follows from the fact that, in regime $R1$, the parent provides her child with a transfer even if his illegal activities remain undetected: $I_P > c^u_{P|R1}$.

By (C.17), $y_1^*(\frac{\phi}{\alpha}s_P) + \hat{\epsilon}_1 < \frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2$: Otherwise, if the parent was indifferent between regimes $R1$ and $R2$ for some $y \geq \frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2$ and the ethic corresponding to interior solutions under regime $R2$, the child would prefer $R1$ over $R2$ instead. Consequently, for $y \in [y_1^*(\frac{\phi}{\alpha}s_P) + \hat{\epsilon}_1, \frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2]$, the parent must choose an ethic $s_C^*(y)$ *weaker* than the corresponding interior-solution level of regime $R2$. As her expected utility associated with $R2$ is inverse-U-shaped with respect to $s_C$, she chooses the largest $s_C^*(y)$ that will keep her child just indifferent between regimes $R1$ and $R2$. Again, by (C.14), $\frac{ds_C^*}{dy} > 0$. Finally, as soon as $s_C^*(y)$ has approached the interior-solution level of regime $R2$, with $y = \tilde{y}_{12}$, the parent strictly prefers $R2$, while her child is indifferent: By (C.17), $V(\cdot|R1) = V(\cdot|R2)$ implies that $E[U_P|R2] > E[U_P|R1]$. Hence, $\tilde{y}_{12} = \frac{\phi}{1-\phi}I_P - \hat{\epsilon}_2$.

For legal income prospects $y > \frac{\phi}{1-\phi}I_P$, the proof follows the above line of reasoning if we let $y$ *decrease* from $\frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P$ down to $\frac{\phi}{1-\phi}I_P$. It suffices to show the following: First, for income prospects $y \geq \tilde{y}_{33}$, the child *weakly* prefers regime $R3$ over $R2$, whereas the parent *strictly* prefers $R3$ over $R2$; second, for income prospects $\tilde{y}_{23}$, the parent prefers regime $R2$ over $R3$; and third, if the parent chooses $s_C^*$ such that $V(\cdot|R2) = V(\cdot|R3)$, then $\frac{ds_C^*}{dy} < 0$.

First, we know from (III.18) and Lemma III.1(v) that, for $y \geq \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P$, equilibrium ethic is given by $s_C^*(y) = \frac{1}{\alpha}s_P$. Furthermore, we know from Lemma III.1 that, for $s_C^* = \frac{1}{\alpha}s_P$, there exists a legal income level $y_2^*(\frac{1}{\alpha}s_P) \in (\frac{\phi}{1-\phi}I_P, \frac{1}{1-\pi}\frac{\phi}{1-\phi}I_P)$ such that the child prefers $R3$ over $R2$ for all $y \geq y_2^*(\frac{1}{\alpha}s_P)$, while changing preferences at $y = y_2^*(\frac{1}{\alpha}s_P)$ if $y$ (hypothetically) decreases. By definition, $\tilde{y}_{33} = y_2^*(\frac{1}{\alpha}s_P)$. We can contrast the child's preferences with the parent's by considering the difference

$$\begin{aligned} E[U_P|R3] - E[U_P|R2] &= p[u(I_P) - u(c^d_{P|R2})] \\ &\quad - p(s_P - \alpha s_C)\pi[g(x_{R2}) - g(x_{R3})] \\ &\quad + \alpha[V(\cdot|R3) - V(\cdot|R2)] \\ &= p[u(I_P) - u(c^d_{P|R2})] \\ &\quad + \alpha[V(\cdot|R3) - V(\cdot|R2)], \end{aligned} \tag{C.18}$$

where we have substituted for $s_C = s_C^* = \frac{1}{\alpha}s_P$. Since $I_P > c^d_{C|R2}$ due to the fact that, in regime $R2$, the parent provides her child with a transfer if his illegal activities are detected, we observe from (C.18) that, as long as the child *weakly* prefers $R3$ over $R2$ as $y$ (hypothetically) decreases, the parent *strictly* prefers $R3$ over $R2$.

Second, consider $\tilde{y}_{23}$ and the corresponding $s_C^*$, such that $V(\cdot|R1) = V(\cdot|R2)$. Taking differences of the parent's expected utility for regimes $R2$ and $R3$, while substituting for $s_C^* = \frac{1}{\alpha}s_P + \frac{1}{\alpha}(1-\phi)u'(c_P^d)$, yields

$$\begin{aligned} E[U_P|R3] - E[U_P|R2] &= p[u(I_P) - u(c^d_{P|R2})] \\ &\quad - pu'(c^d_{P|R2})(1-\phi)\pi[g(x_{R2}) - g(x_{R3})] \\ &\quad + \alpha[V(\cdot|R3) - V(\cdot|R2)]. \end{aligned} \tag{C.19}$$

Concavity of utility from consumption implies that

$$\begin{aligned} u(I_P) - u(c^d_{P|R2}) &< u'(c^d_{P|R2})(I_P - c^d_{P|R2}) \\ &= u'(c^d_{P|R2})[\phi I_P - (1-\phi)I^d_{C|R2}]. \end{aligned} \tag{C.20}$$

On the other hand,

$$\pi g(x_{R2}) - \pi g(x_{R3}) = I^d_{C|R3} - I^d_{C|R2}. \tag{C.21}$$

Substituting (C.20), (C.21), and $V(\cdot|R1) = V(\cdot|R2)$ into (C.19) yields

$$\begin{aligned} E[U_P|R3] - E[U_P|R2] &< pu'(c^d_{P|R2})[\phi I_P - (1-\phi)I^d_{C|R3}] \\ &\leq 0, \end{aligned}$$

where the last inequality follows from the fact that, in regime $R3$, the parent does not provide her child with a transfer, such that $\phi I_P - (1-\phi)I^d_{C|R3} \leq 0$ according to (III.4).

Third, suppose $V(\cdot|R2) = V(\cdot|R3)$. Taking derivatives with respect to $s_C$ and $y$ yields

$$\frac{ds^*_{23}}{dy} = -\frac{V_y(\cdot|R3) - V_y(\cdot|R2)}{V_{s_C}(\cdot|R3) - V_{s_C}(\cdot|R2)}.$$

By (III.8), $V_{s_C}(\cdot|Rj) = -p\, x|_{Rj} < 0$, whereas we know from (III.12) and (III.13) that, *ceteris paribus*, $x_{R2} > x_{R3}$. Therefore, $V_{s_C}(\cdot|R2) < V_{s_C}(\cdot|R3)$. On the other hand, $V_y(\cdot|R3) > V_y(\cdot|R2)$. Hence, $\frac{ds^*_C}{dy} < 0$.

This completes the proof of Proposition III.1. ■

## C.3. Proof of Lemma III.2

The first part of the Lemma is obvious from Equations (III.16) and (III.18) for interior solutions under regimes $R1$ and $R3$.

In regime $R2$, the parent's consumption level if her child is convicted is given by $c^d_{P|R2} = (1-\phi)(I_P + y - \pi g(x_{R2}))$. Implicit differentiation of the parent's first-order condition (III.17) with respect to $p$ yields

$$\frac{\partial s_C}{\partial p} = -\frac{(1-\phi)u''(c^d_{P|R2})\frac{\partial c^d_{P|R2}}{\partial p}}{(1-\phi)u''(c^d_{P|R2})\frac{\partial c^d_{P|R2}}{\partial s_C} - \alpha}. \tag{C.22}$$

As the denominator of the right-hand side of (C.22) is negative due to the parent's second-order condition, the numerator of (C.22) determines the sign of $\frac{\partial s_C}{\partial p}$. Since $u'' < 0$, while $\frac{\partial c^d_{P|R2}}{\partial p} = -(1-\phi)\pi g'(x_{R2})\frac{\partial x_{R2}}{\partial p}$, and $\frac{\partial x_{R2}}{\partial p} < 0$ due to (III.12), the numerator of (C.22) is negative. Hence, an increase in the detection rate weakens the ethics of law abidance. ∎

## C.4. Proof of Lemma III.3

By Proposition III.1, equilibrium ethic is non-interior for legal income prospects $y \in (\tilde{y}_{11}, \tilde{y}_{12})$ and $y \in (\tilde{y}_{23}, \tilde{y}_{33})$. In these cases, the parent sets $s_C$ such that the child is indifferent between regimes $R1$ and $R2$, or between regimes $R2$ and $R3$, respectively: For the child's indirect expected utility in regime $Rj$, $V(s_C, y, p, \pi|Rj) = \max_x E[U_C|Rj]$, equilibrium ethic $s^*_{12} = s^*_C(y)$ for the regime switch from $R1$ to $R2$ and equilibrium ethic $s^*_{23} = s^*_C(y)$ for the regime switch from $R2$ to $R3$ satisfy respectively

$$\begin{aligned} V(s^*_{12}, y, p, \pi|R2) &= V(s^*_{12}, y, p, \pi|R1), \\ V(s^*_{23}, y, p, \pi|R2) &= V(s^*_{23}, y, p, \pi|R3). \end{aligned}$$

Taking derivatives with respect to $\lambda \in \{p, \pi\}$ yields

$$\frac{ds_{12}^*}{d\lambda} = -\frac{V_\lambda(\cdot|R1) - V_\lambda(\cdot|R2)}{V_{s_C}(\cdot|R1) - V_{s_C}(\cdot|R2)}, \tag{C.23}$$

$$\frac{ds_{23}^*}{d\lambda} = -\frac{V_\lambda(\cdot|R3) - V_\lambda(\cdot|R2)}{V_{s_C}(\cdot|R3) - V_{s_C}(\cdot|R2)}. \tag{C.24}$$

We already know from the proof of Proposition III.1 that the denominators of equations (C.23)-(C.24) are positive. Hence, if the numerators are positive (negative), an increase in the parameter $\lambda \in \{p, \pi\}$ of law enforcement weakens (strengthens) ethics formation.

Consider an increase in the detection rate, $p$. Taking derivatives of $V(\cdot|Rj)$ with respect to $p$, while accounting for the fact that the equilibrium crime level $x_{Rj}$ satisfies the respective first-order condition among (III.11)-(III.13), yields

$$V_p(\cdot|R1) = u(c_{C|R1}^d) - u(c_{C|R1}^u) - s_C \pi g(x_{R1}), \tag{C.25}$$

$$V_p(\cdot|R2) = u(c_{C|R2}^d) - u(c_{C|R2}^u) - s_C \pi g(x_{R2}), \tag{C.26}$$

$$V_p(\cdot|R3) = u(c_{C|R3}^d) - u(c_{C|R3}^u) - s_C \pi g(x_{R3}). \tag{C.27}$$

We know that, *ceteris paribus*, the child engages more in illegal activities in regime $R2$ than in either of the other regimes: $x_{R2} > x_{R1}, x_{R3}$. Hence, in regime $R2$, his disposable income if his illegal activities are detected is smaller, and his consumption if his illegal activities remain undetected is larger than in either of the other regimes: $c_{C|R2}^d < c_{C|R1}^d, c_{C|R3}^d$, and $c_{C|R2}^u > c_{C|R1}^u, c_{C|R3}^u$. Comparing (C.25)-(C.27) implies that

$$V_p(\cdot|R1) - V_p(\cdot|R2) > 0,$$

$$V_p(\cdot|R3) - V_p(\cdot|R2) > 0.$$

According to (C.23)-(C.24), this yields $\frac{\partial s_{12}^*}{\partial p}, \frac{\partial s_{23}^*}{\partial p} < 0$. Hence, for non-interior equilibria, an increase in the detection rate weakens the ethic of law abidance.

We show next that $\frac{\partial \tilde{y}_{11}}{\partial p} \geq 0$. Suppose the opposite is true, $\frac{\partial \tilde{y}_{11}}{\partial p} < 0$. Denote by $s_C^* : (\tilde{y}_{11}(p), \tilde{y}_{12}(p)) \to \mathbb{R}$ the equilibrium ethic associated with $p$, and by $\hat{s}_C^* : (\tilde{y}_{11}(\hat{p}), \tilde{y}_{12}(\hat{p})) \to$

$\mathbb{R}$ the equilibrium ethic associated with $\hat{p} > p$. By assumption, $\tilde{y}_{11}(\hat{p}) < \tilde{y}_{11}(p)$. Fix $y' \in (\tilde{y}_{11}(p), \tilde{y}_{12}(p)) \cap (\tilde{y}_{11}(\hat{p}), \tilde{y}_{12}(\hat{p}))$, which exists if $\hat{p} > p$ is chosen sufficiently small. We know from above that $s_C^*(y') > \hat{s}_C^*(y')$. On the other hand, we know from Proposition III.1 that $s_C^*(\tilde{y}_{11}(p)) = \frac{\phi}{\alpha} s_P = \hat{s}_C^*(\tilde{y}_{11}(\hat{p}))$, and that $s_C^*(y)$ and $\hat{s}_C^*(y)$ are continuous and strictly increasing in $y$. Hence, there exists $y'' \in (\tilde{y}_{11}(p), y')$ such that $s_C^*(y'') = \hat{s}_C^*(y'')$, and there exists $y''' \in (\tilde{y}_{11}(p), y'')$ such that $\hat{s}_C^*(y''') > s_C^*(y''')$, which contradicts Lemma III.3. Hence, $\frac{\partial \tilde{y}_{11}}{\partial p} \geq 0$. By a similar line of reasoning, one can show that $\frac{\partial \tilde{y}_{33}}{\partial p} \leq 0$.

Now consider an increase in the fine rate, $\pi$. Notice that the child's indirect expected utility in each regime is given by

$$\begin{aligned} V(\cdot|R1) &= (1-p)u(\phi I_P + \phi(y + f(x_{R1}))) \\ &\quad + pu(\phi I_P + \phi(y - \pi g(x_{R1}))) - p s_C \pi g(x_{R1}), \end{aligned} \tag{C.28}$$

$$\begin{aligned} V(\cdot|R2) &= (1-p)u(y + f(x_{R2})) \\ &\quad + pu(\phi I_P + \phi(y - \pi g(x_{R2}))) - p s_C \pi g(x_{R2}), \end{aligned} \tag{C.29}$$

$$\begin{aligned} V(\cdot|R3) &= (1-p)u(y + f(x_{R3})) \\ &\quad + pu(y - \pi g(x_{R3})) - p s_C \pi g(x_{R3}). \end{aligned} \tag{C.30}$$

Taking derivatives of equations (C.28)-(C.30) with respect to $\pi$, while accounting for the fact that $x_{Rj}$ satisfies the respective first-order condition among (III.11)-(III.13), yields

$$\begin{aligned} V_\pi(\cdot|R1) &= -pg(x_{R1})[\phi p u'(c^d_{C|R1}) + s_C], \\ V_\pi(\cdot|R2) &= -pg(x_{R2})[\phi p u'(c^d_{C|R2}) + s_C], \\ V_\pi(\cdot|R3) &= -pg(x_{R3})[p u'(c^d_{C|R3}) + s_C]. \end{aligned}$$

Hence,

$$V_\pi(\cdot|R1) - V_\pi(\cdot|R2) > 0, \tag{C.31}$$

$$V_\pi(\cdot|R3) - V_\pi(\cdot|R2) \lesseqgtr 0. \tag{C.32}$$

Inequality (C.31) follows from the fact that $x_{R1} < x_{R2}$, and thus $c^d_{C|R1} > c^d_{C|R2}$. The indeterminate sign of (C.32) is caused by the factor $\phi$ in $V_\pi(\cdot|R2)$. For example, when letting $\alpha \downarrow 0$, then $\phi \downarrow 0$, such that regime $R2$ applies only if $I^d_C \downarrow 0$. Since $I^d_C \geq (1-\pi)y$ and $\pi \in (0,1)$, this requires that $y \downarrow 0$ and thus implies that $x_{R2}, x_{R3} \downarrow 0$. In this case, $V_\pi(\cdot|R3) > V_\pi(\cdot|R2)$ if one assumes that $g'(0) > 0$. Hence, $\frac{ds^*_{12}}{d\pi} < 0$ and $\frac{ds^*_{23}}{d\pi} \lesseqgtr 0$. ∎

# List of Figures

# Bibliography


Acemoglu, Daron and Matthew O Jackson. 2015. "History, expectations, and leadership in the evolution of social norms." *Review of Economic Studies* 82 (2):423–456.

———. 2017. "Social norms and the enforcement of laws." *Journal of the European Economic Association* 15 (2):245–295.

Adriani, Fabrizio and Silvia Sonderegger. 2009. "Why do parents socialize their children to behave pro-socially? An information-based theory." *Journal of Public Economics* 93 (11):1119–1124.

Akerlof, George A. 1970. "The market for "lemons": Quality uncertainty and the market mechanism." *Quarterly Journal of Economics* 84 (3):488–500.

Anderson, Elijah. 1999. *Code of the Street.* New York: Norton.

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. "Tax compliance." *Journal of Economic Literature* 36 (2):818–860.

Andreoni, James and John Miller. 2002. "Giving according to GARP: An experimental test of the consistency of preferences for altruism." *Econometrica* 70 (2):737–753.

Arrow, Kenneth J. 1973. "Some ordinalist-utilitarian notes on Rawls's theory of justice." *Journal of Philosophy* 70 (9):245–263.

———. 1979. "The property rights doctrine and demand revelation under incomplete information." In *Economics and Human Welfare*, edited by M. J. Boskin. New York, NY: Academic Press.

Atkinson, Anthony B. 1970. "On the measurement of inequality." *Journal of Economic Theory* 2 (3):244–263.

Ausubel, Lawrence M, Peter Cramton, and Raymond J Deneckere. 2002. "Bargaining with incomplete information." *Handbook of Game Theory with Economic Applications* 3:1897–1945.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2015. "Letting down the team? Social effects of team incentives." *Journal of the European Economic Association* 13 (5):841–870.

Balkundi, Prasad and David A Harrison. 2006. "Ties, leaders, and time in teams: Strong inference about network structures effects on team viability and performance." *Academy of Management Journal* 49 (1):49–68.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2010. "Social incentives in the workplace." *Review of Economic Studies* 77 (2):417–458.

———. 2013. "Team incentives: evidence from a firm level experiment." *Journal of the European Economic Association* 11 (5):1079–1114.

Barber, Brian K, Heidi E Stolz, Joseph A Olsen, W Andrew Collins, and Margaret Burchinal. 2005. "Parental support, psychological control, and behavioral control: Assessing relevance across time, culture, and method." *Monographs of the society for research in child development* 70 (4):i+v+vii+1–147.

Bartling, Björn and Nick Netzer. 2016. "An externality-robust auction: Theory and experimental evidence." *Games and Economic Behavior* 97:186–204.

Becker, Gary S. 1968. "Crime and punishment: An economic approach." *Journal of Political Economy* 76 (2):169–217.

———. 1974. "A Theory of Social Interactions." *Journal of Political Economy* 82 (6):1063–93.

———. 1976. “Altruism, egoism, and genetic fitness: Economics and sociobiology.” *Journal of Economic Literature* 14 (3):817–826.

———. 1981. *A treatise on the family.* Cambridge, MA: Harvard University Press.

Becker, Gary S, Kevin M Murphy, and Jörg L Spenkuch. 2016. “The Manipulation of Childrens Preferences, Old-Age Support, and Investment in Childrens Human Capital.” *Journal of Labor Economics* 34 (S2):S3–S30.

Becker, Gary S and George J Stigler. 1974. “Law enforcement, malfeasance, and compensation of enforcers.” *Journal of Legal Studies* 3 (1):1–18.

Bénabou, Roland and Jean Tirole. 2003. “Intrinsic and extrinsic motivation.” *Review of Economic Studies* 70 (3):489–520.

Bergemann, Dirk and Stephen Morris. 2005. “Robust mechanism design.” *Econometrica* 73 (6):1771–1813.

———. 2009. “Robust implementation in direct mechanisms.” *Review of Economic Studies* 76 (4):1175–1204.

———. 2011. “Robust implementation in general mechanisms.” *Games and Economic Behavior* 71 (2):261–281.

———. 2013. “Robust predictions in games with incomplete information.” *Econometrica* 81 (4):1251–1308.

Bester, Helmut and Daniel Krähmer. 2008. “Delegation and incentives.” *RAND Journal of Economics* 39 (3):664–682.

Bidner, Chris and Patrick Francois. 2010. “Cultivating Trust: Norms, Institutions and the Implications of Scale.” *Economic Journal* 121 (6):1097–1129.

Bierbrauer, Felix and Nick Netzer. 2016. “Mechanism design and intentions.” *Journal of Economic Theory* 163:557–603.

Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert. 2017. "Robust mechanism design and social preferences." *Journal of Public Economics* 149:59–80.

Bisin, Alberto and Thierry Verdier. 1998. "On the cultural transmission of preferences for social status." *Journal of Public Economics* 70 (1):75–97.

———. 2001. "The economics of cultural transmission and the dynamics of preferences." *Journal of Economic Theory* 97 (2):298–319.

———. 2010. "The economics of cultural transmission and socialization." NBER Working Paper No. 16512.

Boskin, Michael J and Eytan Sheshinski. 1978. "Optimal redistributive taxation when individual welfare depends upon relative income." *Quarterly Journal of Economics* 92 (4):589–601.

Bowles, Samuel and Sung-Ha Hwang. 2008. "Social preferences and public economics: Mechanism design when social preferences depend on incentives." *Journal of Public Economics* 92 (8):1811–1820.

Bramoullé, Yann and Rachel Kranton. 2007. "Public goods in networks." *Journal of Economic Theory* 135 (1):478–494.

Bruce, Neil and Michael Waldman. 1990. "The rotten-kid theorem meets the Samaritan's dilemma." *Quarterly Journal of Economics* 105 (1):155–165.

Brunner, Markus and Kai Jörg Sandner. 2012. "Social comparison, group composition, and incentive provision." *International Journal of Game Theory* 41 (3):565–602.

Buchanan, James M. 1975. "The Samaritan's dilemma." In *Altruism, morality and economic theory*, edited by E.S. Phelps. New York: Russel Sage Foundation, 71–85.

Cabrales, Antonio and Roberto Serrano. 2011. "Implementation in adaptive better-response dynamics: Towards a general theory of bounded rationality in mechanisms." *Games and Economic Behavior* 73 (2):360–374.

Carpenter, Jeffrey and Erika Seki. 2011. "Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay." *Economic Inquiry* 49 (2):612–630.

Charness, Gary, Ramón Cobo-Reyes, Natalia Jiménez, Juan A Lacomba, and Francisco Lagos. 2012. "The hidden advantage of delegation: Pareto improvements in a gift exchange game." *American Economic Review* 102 (5):2358–2379.

Charness, Gary, David Masclet, and Marie Claire Villeval. 2013. "The dark side of competition for status." *Management Science* 60 (1):38–55.

Charnock, E. 2010. *E-Habits: What You Must Do to Optimize Your Professional Digital Presence*. McGraw-Hill Education.

Chetty, Raj and Nathaniel Hendren. 2016a. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Eects." NBER Working Paper No. 23001.

———. 2016b. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." NBER Working Paper No. 23002.

Chung, Kim-Sau and Jeffrey C Ely. 2007. "Foundations of dominant-strategy mechanisms." *Review of Economic Studies* 74 (2):447–476.

Clark, Andrew E, Paul Frijters, and Michael A Shields. 2008. "Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles." *Journal of Economic Literature* 46 (1):95–144.

Clarke, Edward H. 1971. "Multipart pricing of public goods." *Public Choice* 11 (1):17–33.

Dasgupta, Partha and Eric Maskin. 2000. "Efficient auctions." *Quarterly Journal of Economics* 115 (2):341–388.

d'Aspremont, Claude and Louis-André Gérard-Varet. 1979. "Incentives and incomplete information." *Journal of Public Economics* 11 (1):25–45.

d'Aspremont, Claude and Louis Gevers. 1977. "Equity and the informational basis of collective choice." *Review of Economic Studies* 44 (2):199–209.

de Clippel, Geoffroy. 2014. "Behavioral implementation." *American Economic Review* 104 (10):2975–3002.

Dessi, Roberta. 2008. "Collective memory, cultural transmission, and investments." *American Economic Review* 98 (1):534–560.

Doepke, Matthias and Fabrizio Zilibotti. 2008. "Occupational choice and the spirit of capitalism." *Quarterly Journal of Economics* 123 (2):747–793.

———. 2017. "Parenting with style: Altruism and paternalism in intergenerational preference transmission." *Econometrica* 85 (5):1331–1371.

Dutta, Bhaskar, Sayantan Ghosal, and Debraj Ray. 2005. "Farsighted network formation." *Journal of Economic Theory* 122 (2):143–164.

Easterlin, Richard A. 1974. "Does economic growth improve the human lot? Some empirical evidence." In *Nations and Households in Economic Growth*, edited by P. A. David and M. W. Reder. New York, NY: Academic Press, 89–125.

Elfenbein, Hillary Anger and Charles A O'Reilly. 2007. "Fitting in: The effects of relational demography and person-culture fit on group process and performance." *Group & Organization Management* 32 (1):109–142.

Ellickson, Robert C. 1998. "Law and economics discovers social norms." *Journal of Legal Studies* 27 (S2):537–552.

Falk, Armin and Michael Kosfeld. 2006. "The hidden costs of control." *American Economic Review* 96 (5):1611–1630.

Farrington, David P. 2003. "Key results from the first forty years of the Cambridge study in delinquent development." In *Taking stock of delinquency*. Springer, 137–183.

Fehr, Ernst and Simon Gächter. 2002. "Altruistic punishment in humans." *Nature* 415 (6868):137–140.

Fehr, Ernst and Klaus M Schmidt. 2006. "The economics of fairness, reciprocity and altruism–experimental evidence and new theories." *Handbook of the Economics of Giving, Altruism and Reciprocity* 1:615–691.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are people conditionally cooperative? Evidence from a public goods experiment." *Economics Letters* 71 (3):397–404.

Flores-Fillol, Ricardo, Susana Iranzo, and Ferran Mane. 2017. "Teamwork and delegation of decisions within the firm." *International Journal of Industrial Organization* 52:1–29.

Frank, Robert H. 1984. "Are workers paid their marginal products?" *American Economic Review* 74 (4):549–571.

———. 1985. *Choosing the right pond: Human behavior and the quest for status.* Oxford University Press.

Frey, Bruno S and Reto Jegen. 2001. "Motivation crowding theory." *Journal of Economic Surveys* 15 (5):589–611.

Galeotti, Andrea, Sanjeev Goyal, Matthew O Jackson, Fernando Vega-Redondo, and Leeat Yariv. 2010. "Network games." *Review of Economic Studies* 77 (1):218–244.

Gershkov, Alex, Jacob K Goeree, Alexey Kushnir, Benny Moldovanu, and Xianwen Shi. 2013. "On the equivalence of Bayesian and dominant strategy implementation." *Econometrica* 81 (1):197–220.

Gershkov, Alex, Jianpei Li, and Paul Schweinzer. 2009. "Efficient tournaments within teams." *RAND Journal of Economics* 40 (1):103–119.

Glazer, Jacob and Ariel Rubinstein. 1998. "Motives and implementation: On the design of mechanisms to elicit opinions." *Journal of Economic Theory* 79 (2):157–173.

Goeree, Jacob K, Emiel Maasland, Sander Onderstal, and John L Turner. 2005. "How (not) to raise money." *Journal of Political Economy* 113 (4):897–918.

Goette, Lorenz, David Huffman, Stephan Meier, and Matthias Sutter. 2012. "Competition between organizational groups: Its impact on altruistic and antisocial motivations." *Management Science* 58 (5):948–960.

Gordon, James PP. 1989. "Individual morality and reputation costs as deterrents to tax evasion." *European Economic Review* 33 (4):797–805.

Groves, Theodore. 1973. "Incentives in teams." *Econometrica* 41 (4):617–631.

Groves, Theodore and John Ledyard. 1977. "Optimal allocation of public goods: A solution to the" free rider" problem." *Econometrica* 45 (4):783–809.

Hamilton, Barton H, Jack A Nickerson, and Hideo Owan. 2003. "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation." *Journal of Political Economy* 111 (3):465–497.

Harsanyi, John C. 1955. "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility." *Journal of Political Economy* 63 (4):309–321.

Harsanyi, John C and Reinhard Selten. 1972. "A generalized Nash solution for two-person bargaining games with incomplete information." *Management Science* 18 (5-part-2):80–106.

Harstad, Bård. 2007. "Harmonization and side payments in political cooperation." *American Economic Review* 97 (3):871–889.

Hauk, Esther and Maria Saez-Martib. 2002. "On the Cultural Transmission of Corruption." *Journal of Economic Theory* 107 (2):311–335.

Heckman, James J. and Stefano Mosso. 2014. "The economics of human development and social mobility." *Working Paper 19925, National Bureau of Economic Research* .

Heffetz, Ori and Robert H Frank. 2008. "Preferences for status: Evidence and economic implications." In *Handbook of Social Economics*, vol. 1, edited by Jess Benhabib, Alberto Bisin, and Matthew Jackson. Netherlands: North Holland, 69–91.

Hirschman, Albert O and Michael Rothschild. 1973. "The changing tolerance for income inequality in the course of economic development with a mathematical appendix." *Quarterly Journal of Economics* 87 (4):544–566.

Hoffmann, L. 2010. "Mind your business." *Communications of the ACM* 53 (6):18–19.

Holmstrom, Bengt. 1982. "Moral hazard in teams." *Bell Journal of Economics* 13 (2):324–340.

Huckman, Robert S, Bradley R Staats, and David M Upton. 2009. "Team familiarity, role experience, and performance: Evidence from Indian software services." *Management Science* 55 (1):85–100.

Ingvaldsen, Jonas A and Monica Rolfsen. 2012. "Autonomous work groups and the challenge of inter-group coordination." *Human Relations* 65 (7):861–881.

Itoh, Hideshi. 1991. "Incentives to help in multi-agent situations." *Econometrica* 59 (3):611–636.

Jackson, Matthew O and Asher Wolinsky. 1996. "A strategic model of social and economic networks." *Journal of Economic Theory* 71 (1):44–74.

Jehiel, Philippe, Moritz Meyer-ter Vehn, Benny Moldovanu, and William R Zame. 2006. "The limits of ex post implementation." *Econometrica* 74 (3):585–610.

Jehiel, Philippe and Benny Moldovanu. 2001. "Efficient design with interdependent valuations." *Econometrica* 69 (5):1237–1259.

Jehiel, Philippe, Benny Moldovanu, and Ennio Stacchetti. 1996. "How (not) to sell nuclear weapons." *American Economic Review* 86 (4):814–829.

———. 1999. "Multidimensional mechanism design for auctions with externalities." *Journal of Economic Theory* 85 (2):258–293.

Kalai, Ehud. 1977. "Proportional solutions to bargaining situations: interpersonal utility comparisons." *Econometrica* 45 (7):1623–1630.

Kalai, Ehud and Meir Smorodinsky. 1975. "Other solutions to Nash's bargaining problem." *Econometrica* 43 (3):513–518.

Kandel, Eugene and Edward P Lazear. 1992. "Peer pressure and partnerships." *Journal of Political Economy* 100 (4):801–817.

Kreps, David M. 1997. "Intrinsic motivation and extrinsic incentives." *American Economic Review: Papers and Proceedings* 87 (2):359–364.

Kucuksenel, Serkan. 2012. "Behavioral mechanism design." *Journal of Public Economic Theory* 14 (5):767–789.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics* 1 (3):593–622.

Lindbeck, Assar and Sten Nyberg. 2006. "Raising children to work hard: altruism, work norms, and social insurance." *Quarterly Journal of Economics* 121 (4):1473–1503.

Lu, Jingfeng. 2012. "Optimal auctions with asymmetric financial externalities." *Games and Economic Behavior* 74 (2):561–575.

Luttmer, Erzo FP and Monica Singhal. 2014. "Tax Morale." *Journal of Economic Perspectives* 28 (4):149–68.

MacDonald, Robert and Jane Marsh. 2005. *Disconnected Youth? Growing up in Britians Poor Neighborhood.* Palgrave MacMillian.

Mas, Alexandre and Enrico Moretti. 2009. "Peers at work." *American Economic Review* 99 (1):112–145.

Mas-Colell, Andreu, Michael D Whinston, and Jerry R Green. 1995. *Microeconomic theory.* New York: Oxford University Press.

Maskin, Eric. 1978. "A theorem on utilitarianism." *Review of Economic Studies* 45 (1):93–96.

———. 1979. “Decision-making under ignorance with implications for social choice.” In *Game Theory, Social Choice and Ethics*. Springer, 319–337.

———. 1999. “Nash equilibrium and welfare optimality.” *Review of Economic Studies* 66 (1):23–38.

McAdams, Richard H and Eric B Rasmusen. 2007. “Norms and the law.” In *Handbook of Law and Economics*, vol. 2, edited by A. Mitchell Polinsky and Steven Shavell. Amsterdam: Elsevier, 1573–1618.

McAfee, R Preston and John McMillan. 1991. “Optimal contracts for teams.” *International Economic Review* 32 (3):561–577.

McCord, Joan. 1991. “Family relationships, juvenile delinquency, and adult criminality.” *Criminology* 29 (3):397–417.

McFadden, Daniel. 2009. “The human side of mechanism design: a tribute to Leo Hurwicz and Jean-Jacque Laffont.” *Review of Economic Design* 13 (1-2):77–100.

Mirrlees, James A. 1971. “An exploration in the theory of optimum income taxation.” *Review of Economic Studies* 38 (2):175–208.

Moldoveanu, Mihnea C and Joel AC Baum. 2011. “I Think You Think I Think You’re Lying: The Interactive Epistemology of Trust in Social Networks.” *Management Science* 57 (2):393–412.

Mookherjee, Dilip and Stefan Reichelstein. 1992. “Dominant strategy implementation of Bayesian incentive compatible allocation rules.” *Journal of Economic Theory* 56 (2):378–399.

Moore, John and Rafael Repullo. 1990. “Nash implementation: a full characterization.” *Econometrica* 58 (5):1083–1099.

Myerson, Roger B. 1979. “Incentive compatibility and the bargaining problem.” *Econometrica* 47 (1):61–73.

Myerson, Roger B and Mark A Satterthwaite. 1983. "Efficient mechanisms for bilateral trading." *Journal of Economic Theory* 29 (2):265–281.

Nash, John F. 1950. "The bargaining problem." *Econometrica* 18 (2):155–162.

Neuman, George A and Julie Wright. 1999. "Team effectiveness: Beyond skills and cognitive ability." *Journal of Applied Psychology* 84 (3):376–389.

Oh, Hongseok, Myung-Ho Chung, and Giuseppe Labianca. 2004. "Group social capital and group effectiveness: The role of informal socializing ties." *Academy of Management Journal* 47 (6):860–875.

Page, Frank H, Myrna H Wooders, and Samir Kamat. 2005. "Networks and farsighted stability." *Journal of Economic Theory* 120 (2):257–269.

Polinsky, A Mitchell and Steven Shavell. 2000. "The economic theory of public enforcement of law." *Journal of Economic Literature* 38 (1):45–76.

Posner, Richard A. 1979. "Utilitarianism, Economics, and Legal Theory." *Journal of Legal Studies* 8 (1):103–140.

———. 1997. "Social norms and the law: An economic approach." *American Economic Review: Papers and Proceedings* 87 (2):365–369.

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." *American Economic Review* 83 (5):1281–1302.

Rawls, John B. 1971. *A theory of justice.* Harvard University Press.

Reagans, Ray, Ezra Zuckerman, and Bill McEvily. 2004. "How to make the team: Social networks vs. demography as criteria for designing effective teams." *Administrative Science Quarterly* 49 (1):101–133.

Rotemberg, Julio J. 1994. "Human relations in the workplace." *Journal of Political Economy* 102 (4):684–717.

Rothschild, Michael and Joseph Stiglitz. 1976. "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information." *Quarterly Journal of Economics* 90 (4):629–649.

Rubinstein, Ariel. 1982. "Perfect equilibrium in a bargaining model." *Econometrica* 50 (1):97–109.

Rulke, Diane L and Joseph Galaskiewicz. 2000. "Distribution of knowledge, group network structure, and group performance." *Management Science* 46 (5):612–625.

Saez, Emmanuel and Stefanie Stantcheva. 2016. "Generalized social marginal welfare weights for optimal tax theory." *American Economic Review* 106 (1):24–45.

Saijo, Tatsuyoshi and Hideki Nakamura. 1995. "The spite dilemma in voluntary contribution mechanism experiments." *Journal of Conflict Resolution* 39 (3):535–560.

Sen, Amartya. 1973. *On economic inequality.* Oxford University Press.

———. 1974. "Informational bases of alternative welfare approaches: aggregation and income distribution." *Journal of Public Economics* 3 (4):387–403.

———. 1979. "Utilitarianism and welfarism." *Journal of Philosophy* 76 (9):463–489.

Spence, Michael. 1973. "Job market signaling." *Quarterly Journal of Economics* 87 (3):355–374.

Tabellini, Guido. 2008. "The scope of cooperation: values and incentives." *Quarterly Journal of Economics* 123 (3):905–950.

Tang, Pingzhong and Tuomas Sandholm. 2012. "Optimal Auctions for Spiteful Bidders." In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. pp. 1457–1463.

Thomson, William. 1994. "Cooperative models of bargaining." *Handbook of Game Theory with Economic Applications* 2:1237–1284.

Tran, Anh and Richard Zeckhauser. 2012. "Rank as an inherent incentive: Evidence from a field experiment." *Journal of Public Economics* 96 (9):645–650.

Van Dijk, Frans, Joep Sonnemans, and Frans van Winden. 2002. "Social ties in a public good experiment." *Journal of Public Economics* 85 (2):275–299.

Vickrey, William. 1961. "Counterspeculation, auctions, and competitive sealed tenders." *Journal of Finance* 16 (1):8–37.

Weinberg, Bruce A. 2001. "An incentive model of the effect of parental income on children." *Journal of Political Economy* 109 (2):266–280.

Weiss, Yoram and Chaim Fershtman. 1998. "Social status and economic performance: A survey." *European Economic Review* 42 (3):801–820.

Wilson, Robert. 1987. "Game-theoretic analyses of trading processes." In *Advances in Economic Theory: Fifth World Congress*, edited by T. Bewley, chap. 2. Cambridge U.K.: Cambridge University Press, 33–70.

Wilson, William J. 1997. *When Work Disappears: The World of the New Urban Poor*. New York: Alfred A. Knopf.

Young, Tara, Wendy Fitzgibbon, and Daniel Silverstone. 2014. "A question of family? Youth and gangs." *Youth Justice* 14 (2):171–185.