Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2015/3

The Expected Externality Mechanism in a Level-*k* Environment

Olga Gorelkina

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2015/3

The Expected Externality Mechanism in a Level-*k* Environment

Olga Gorelkina

January 2015

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

The Expected Externality Mechanism in a Level-k Environment*

Olga Gorelkina†

Abstract

Mechanism design theory strongly relies on the concept of Nash equilibrium. However, studies of experimental games show that Nash equilibria are rarely played and that subjects may be thinking only a finite number of iterations. We study one of the most influential benchmarks of mechanism design theory, the expected externality mechanism (D'Aspremont, Gerard-Varet, 1979) in a finite-depth environment described by the *Lk* model. While efficient implementation fails under certain conditions, our results provide a vindication of the mechanism in the convex quasi-linear environment with finitely-rational agents.

1 Introduction

Mechanism design theory studies institutions with privately informed agents. Using the tools of game theory, it proposes rules of interactions such that the participants' strategic behavior complies with the designer's objective. In a leading example, the designer's purpose is to implement the socially efficient outcome, that is, to find the allocation that maximizes total welfare. The major

*I am grateful to Vincent Crawford, Francoise Forges, Ioanna Grypari, Rida Laraki, Thomas Mariotti, David Martimort, Benny Moldovanu, Thomas Rieck, Nicolas Roux for their helpful comments.

†Max Planck Institute for Research on Collective Goods, Bonn, Germany.

challenge to efficient implementation is the fact that information about individual preferences is private.[1] In a setting with quasi-linear utilities, d'Aspremont and Gerard-Varet (AGV, 1979) construct an ingenious mechanism that aligns the agents' individual incentives with total welfare maximization. In a Bayes-Nash equilibrium, the agents' then reveal their types truthfully and efficiency is achieved. The AGV mechanism has become an essential building block for the mechanism design theory (Athey and Segal, 2013).

Since the AGV mechanism is tailored to the concept of Bayes-Nash equilibrium, its success in inducing truth-telling and, therefore, efficiency in practice depends on (i) whether the participants' behavioral response to the mechanism coincides with the Bayes-Nash prediction and, if it does not, (ii) whether efficiency still obtains under the possible deviations. While the first question has not been addressed directly in the literature, the experimental results in (simpler) complete information games suggest that the answer is negative. As to the second question, little is known as to the loss of efficiency if the participants do *not* play equilibrium.[2] This paper tries to fill this gap by studying how the mechanism performs in a behavioral framework where, contrary to the requirement of Bayes-Nash equilibrium, the agents conduct only a limited number of iterations of reasoning.

The choice of behavioral setting follows a large body of evidence from experimental games. Recent surveys by Crawford, Costa-Gomez, and Iriberri (2013) and Camerer and Ho (2015) show that non-equilibrium models with finite depth of reasoning, such as the *Level-k* model (*Lk*; Nagel 1995; Stahl and Wilson 1994; Costa-Gomes, Crawford, and Broseta 2001; Costa-Gomes and Crawford 2006) and the *cognitive hierarchy* model (CH; Camerer, Ho, and Chong 2004), systematically outperform equilibrium in predicting human behavior. Along with closely fitting the lab data, these models are able to predict some frequently observed field phenomena such as the winner's curse in common-value auctions: see Crawford and Iriberri, 2007. We choose the *Lk* model due to its tractability, but most of our results also hold in the CH model.[3]

Lk is a model of reasoning prior to a game, where the player maximizes his payoff against a non-equilibrium belief about other players' strategies. The belief

[1] In this literature, all private information is summarized in a *type:* a parameter that enters the player's utility function (and has to be elicited by the mechanism).

[2] See Crawford, Costa-Gomez, and Iriberri (2013).

[3] Propositions 1, 2, 4, and 5 hold in the cognitive hierarchy model.

is constructed in the following iterative process. A player of level $k = 1$ (*"L1* player*"*) believes that his opponents (*"L0"*) behave non-strategically. In incomplete information games, such as the AGV mechanism, *L0*'s can be modeled in two distinct ways: either they truthfully reveal their type ("truthful *L0*") or draw their actions (type reports) from a random distribution ("random *L0*"). An *L2* player best replies to the profile of *L1* strategies, *L3* best replies to *L2*, and so on. In general, an *Lk* strategy is best reply to the profile of *Lk-1*, suggesting the interpretation that players try to "outguess" their opponents.[4] As an illustration, think of a game, where the players pick a number between 0 and 100 and the one whose number is closest to some fraction, say one half, of the average wins the game. In this guessing game, if *L0*s randomize uniformly between 0 and 100, *L1*s will choose 50/2=25, *L2*s will choose 25/2, etc. As *Lk* increases, the best response approaches 0, the only Nash equilibrium of the game.

We apply the *Lk* model to the AGV mechanism in a setting with independent private valuations and utilities that are strictly concave with respect to the allocation.[5] First, we observe that in the truthful-*L0* specification of the *Lk* model the mechanism never produces a loss in efficiency. In that specification, the *L1* best reply is given by the equilibrium condition of AGV which implies truth-telling. By induction, this result extends to any higher level *k*, therefore the mechanism chooses the efficient allocation irrespectively the levels prevailing in the population.

Further, in the random-*L0* specification of *Lk*, we show that if the distribution of random moves (*L0*) coincides with the distribution of payoff types, then the participants at any level larger than zero report truthfully to the mechanism. Next, we analyze the more challenging setup where the type distribution used by the planner to assign transfers differs from *L1*s' perception of the opponents' moves. In this case, the externality payment generally fails to align the agent's incentives with total expected welfare maximization. As a result, the AGV mechanism does not induce truth-telling and produce a suboptimal allocation. Denoting the distribution of random *L0* strategies by Φ and the distribution of types by F, we study how the stochastic properties of Φ and F affect the *Lk* strategies in the mechanism.

[4]The cognitive hierarchy model features 'smoother' beliefs: a positive probability is assigned to *all* levels lower than one's own.

[5]We use the assumption of strict concavity to assure that the equilibrium of the AGV mechanism is unique. For an account of the problem of non-uniqueness, see Mathevet (2010).

We start with a simple environment where utilities are quadratic. In this setting, a difference in the mean values of Φ and F creates distortions at level 1. For instance, if the mean type is greater than the mean *L0* report, then all types of an *L1* player will over-report their types to the mechanism. Misreporting carries over to higher levels, but the expected absolute value of the distortion of type decreases exponentially as level *k* goes up. Moreover, the direction of bias (i.e., whether the agents over-report or under-report their types) alternates at each iteration from *k* to *k+1*. This result has two interesting implications for the outcome of the mechanism. First, if the pool of agents is a mixture of two subsequent levels (e.g., *L2* and *L3*), the distortion of efficiency is lower than in a group where only one of these levels is present. Second, as *Lk* goes up, the outcome approaches efficiency.

Similar results are obtained in a more general setting, where the efficient rule is essentially linear in types.[6] In this neutral environment types are neither substitutes nor complements with respect to the optimal allocation. A simple example of a neutral environment is the one where the optimal allocation is a linear combination, for instance, the average, of types. In this environment, whenever Φ (F) dominates F (Φ) in the sense of first-order stochastic dominance then types are going to be systematically misreported.[7] We find that if the distribution of types F dominates the distribution of random moves Φ, then *L1*s always over-report their types. Thus they compensate the downward bias of *L0*s distorting their own reports in the opposite direction. This is due to the incentive scheme induced by the mechanism: it punishes for the expected, as opposed to the realized, negative externality.

As an extension, we study the case where type reports are complements or substitutes with respect to the optimal allocation. The direction of bias in reports can be predicted, similarly to the neutral case, but only for a subset of types. For instance, one of the results states that if the distribution of types F dominates the distribution of random moves Φ and type reports are complements with respect to the social choice function, then low-type *L1*s over-report their types. The reason that high-type *L0*s will not necessarily do so is that over-reporting leads, in expectation, to an excessively high allocation due to the complementarity in

[6] By 'essentially linear in types' I mean linear in some strictly monotone functions of types.

[7] The stochastic dominance relation corresponds to a biased perception of opponents' behavior, which can be caused by previous experience of play or probability weighting. See Kahneman and Tversky, 1981.

agents' reports.

In the neutral case, we also obtain the following convergence result: as level k increases, the players' strategies in the AGV mechanism tend to truth-telling. Since in most experiments the estimated values of k rarely exceed 3 (Crawford, Costa-Gomez, and Iriberri, 2013; Camerer, Ho, 2015), the convergence result bears little importance for one-shot mechanisms. However, with the interpretation of Lk model as a learning algorithm, this result has an important implication for mechanisms that are played repeatedly.[8] We describe a learning algorithm in the game of incomplete information with a large number of players that is equivalent to the Lk model. If learning follows that algorithm, then our convergence result for Lk implies that the players will gradually learn to report types truthfully. We can interpret the results as a vindication of the AGV mechanism in a convex quasi-linear environment with independent private values. The analysis shows that even if agents are finitely-rational, their behavior in the mechanism is centered around truth-telling.[9]

The rest of this paper is organized as follows. Section 2 presents the key assumptions, the Lk model in incomplete information games and in the AGV mechanism in particular. Section 3 describes the properties of Lk strategies in the AGV mechanism: equivalence of Lk and equilibrium models in the AGV mechanism, the biases due to first order stochastic dominance and convergence in the neutral environment. Section 4 partially extends the results to the case when types are substitutes or complements with respect to the efficient allocation. Section 5 explains how the results can be understood in the context of a learning model, and finally, Section 6 discuses the implications for the practical implementation of the AGV mechanism.

2 The Model

Preferences The preference environment is characterized by the following assumptions:

$A1$. Utilities are linear in money.

[8] This follows the paradigm of evolutionary game theory that suggests a different view, relative to the models of reasoning, on learning to play a game. See Sandholm (2010), Hofbauer and Sandholm (2002), Dekel, Fudenberg, Levine (2004).

[9] This is not the case, for example, in first price auctions. See Crawford, Iriberri (2007).

*A*2. Values are private.

*A*3. Values are independent and identically distributed.

Assumptions *A*1 and *A*2 imply that the utility function of a given agent $i \in I$ can be represented as:

$$v_i(x, \theta_i) + m_i, \tag{1}$$

where $v_i(x, \theta_i)$ is the utility derived from allocation $x \in X$, $\theta_i \in \Theta \subseteq \mathcal{R}$ is the privately known preference parameter that we refer to as the player's *type*, and m_i is the monetary transfer to player i. We assume that $v_i(x, \theta_i)$ is strictly concave in x and continuously differentiable with respect to both arguments. *A*3 implies that the values θ_i are drawn independently across $i \in I$. We denote the respective cumulative function F and assume that F is common knowledge. We require that the preferences satisfy a single crossing (Spence-Mirrlees) condition. The condition postulates that function $v_i(x, \theta_i)$ has a cross-derivative with respect to allocation x and type θ_i with a sign that is constant over the function's domain:

*A*4. $v_i(x, \theta_i)$ satisfies the Spence-Mirrlees condition, i.e., either *A*4.1 or *A*4.2 holds:

*A*4.1 $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x, \theta_i) > 0$, for all i and $(x, \theta_i) \in (X, \Theta)$,

*A*4.2 $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x, \theta_i) < 0$, for all i and $(x, \theta_i) \in (X, \Theta)$.

*A*1-*A*4 are the basic assumptions of mechanism design. A further standard assumption is the common knowledge of rationality: the knowledge that the opponent is rational, the knowledge that the opponent knows that his opponent is rational, and so on *ad infinitum*. In this paper, we consider the case with a finite number of rationality iterations. This frame of reasoning is described by the following model (Nagel, 1995; Crawford and Iriberri, 2007).

Level-*k* Consider a game of incomplete information where the payoffs are given by $u_i(s; \theta_i)$, for each player $i \in I$ of type θ_i and strategy profile $s = \left(s_1, s_2, ..s_{|I|}\right)$, where $s_i \in S$. (We use s_i and $s_i(\theta_i)$ interchangeably.) We look at players who engage in iterations of best reply, following Nagel (1995). The *Lk* strategy $s_i^{(k)}(\theta_i)$ is recursively defined as the function of the player's type θ_i that maximizes his

expected payoff against level-$(k-1)$ profile $s_{-i}^{(k-1)}(\theta_{-i})$.[10] As the starting point of recursion, the model features nonstrategic *L0* players, that can be modeled in two alternative ways (see Crawford, Costa-Gomez, and Iriberri, 2013). In one specification, the *L0*'s always reveal their type truthfully; in the other, *L0*'s actions are drawn from a random distribution. In the version with random *L0*'s, we denote the associated cumulative distribution function by Φ and assume, as it is standard in the *Lk* model, that Φ is known.[11] The formal definition is then the following.

Definition For $k \geq 1$ the optimal strategy $s_i^{(k)}$ maximizes the expected payoff of player i against $s_{-i}^{(k-1)}$:[12]

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S} \mathbb{E}\left[u_i\left(s_i, s_{-i}^{(k-1)}(\theta_{-i}); \theta_i\right)\right], \tag{2}$$

where θ_{-i} is the residual profile of types. For $k = 0$, action $s_i^{(0)}(\theta_i) = s_i^{(0)} \in S$ is a random draw from Φ.

The following simple lemma establishes the connection between the *Lk* and equilibrium strategy profiles.

Lemma 1 If $s^{(k)}(\theta) = s^{(k+1)}(\theta)$ for all $k \geq 1$ and $\theta \in \Theta$, then $s^{(k)}$ is a Bayes-Nash equilibrium.

The lemma follows immediately from Equation (2) and the Bayes-Nash equilibrium conditions: The strategy profile $s^{(k)}(\theta)$ that satisfies the condition of Lemma 1 is a fixed point of best-reply correspondence (2).

Choice Rules and Mechanisms For a quasilinear utility representation (1), we define a choice rule $x^*(\theta)$ as *efficient* if it maximizes the total welfare for every profile of agents' types $\theta = \left(\theta_1, \theta_2, ..\theta_{|I|}\right)$:[13]

[10] In other words, the player believes with certainty that the opponents make exactly $k-1$ iterations of best reply. In contrast, the cognitive hierarchy model assumes that an *Lk* player attributes strictly positive probabilities to *all* the levels of rationality lower than k.

[11] Otherwise the optimal *Lk* strategies are not well-defined.

[12] Assuming strict concavity of the payoff functions.

[13] We restrict the attention to strictly convex problems, such that for all $\theta \in \Theta^{|I|}$ the solution $x^*(\theta)$ to the welfare maximization problem is unique.

$$x^{*}(\theta) = \arg\max_{x \in X} \sum_{i} v_i(x; \theta_i) \tag{3}$$

A (direct) *mechanism* is a system of communication and decision-making, where the privately informed agents report their payoff types and the central authority assigns the allocation and transfers based on the submitted reports. Formally, it is a tuple $\left(x(s), T_1(s), T_2(s), ..T_{|I|}(s)\right)$ of allocation and transfers, such that the payoffs in the mechanism are given by:

$$u_i = v_i(x(s), \theta_i) + T_i(s). \tag{4}$$

A mechanism *implements* choice rule $x(s)$ if the profile of truth-telling strategies, $s_i = \theta_i$, $\forall i$, is an equilibrium.

Expected Externality Mechanism The expected externality mechanism introduced in d'Aspremont and Gerard-Varet (AGV, 1979) implements the efficient allocation in a Bayes-Nash equilibrium. In this mechanism, the center chooses the allocation $x^{*}(s)$ defined in (3) and assigns the following monetary transfers to the participants:

$$T_i(s) = t_i(s_i) - \frac{1}{|I| - 1} \sum_{l \neq i} t_l(s_l), \tag{5}$$

where

$$t_i(s_i) = \mathbb{E} \sum_{j \neq i} v_j\left(x^{*}(s_i, \theta_{-i}); \theta_j\right). \tag{6}$$

The transfer T_i is constructed such that agent i internalizes the expected effect of his report on others. The incentive part of t_i represents the monetary value of externality imposed by the agent's report s_i on others' welfare; the externality is evaluated under the assumption that the other agents report their types truthfully. Therefore, if the agent also expects others to report truthfully (the equilibrium assumption), then his incentives are aligned with total welfare maximization, and there is no benefit in misrepresenting his own true preferences. Thus, in the Bayesian setting, the transfer induces truth-telling in equilibrium (d'Aspremont and Gerard-Varet, 1979). Their result immediately implies that in

the truthful-*L0* specification of the *Lk* model efficient implementation obtains for any k.

The second part of the transfer, $\frac{1}{|I|-1}\sum_{l\neq i} t_l(s_l)$, guarantees that mechanism satisfies ex post budget balance, i.e., its transfers sum up to zero for any profile of reports s (and, in particular, in the level-k model.)[14] Observe that the budget-balancing term $\frac{1}{|I|-1}\sum_{l\neq i} t_l(s_l)$ does not depend on agent i's own report s_i. Therefore this term does not affect best replies and can be omitted in the *Lk* analysis.

Level-*k* in the Mechanism In the expected externality mechanism, a *Lk* player, $k \geq 1$, maximizes the expected gain in the mechanism:

$$\mathbb{E}\left[v_i\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right);\theta_i\right) + t_i(s_i)\right] \tag{7}$$

Given the incentive transfer (6), the optimal *Lk* strategy in the mechanism is defined by the following:[15]

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i\in\Theta} \mathbb{E}\left[v_i\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right);\theta_i\right) + \sum_{j\neq i} v_j\left(x^*(s_i,\theta_{-i});\theta_j\right)\right] \tag{8}$$

By Lemma 1, a strategy profile that satisfies $s^{(k)}(\theta) = s^{(k-1)}(\theta)$ for all k and θ is a Bayes-Nash equilibrium. In particular, if truth-telling obtains at all levels up to $k-1$, for all i and θ_i, then we can substitute $s_i^{(k-1)}(\theta_i) = \theta_i$ in Equation (8) and obtain the equilibrium condition:

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i\in\Theta} \mathbb{E}\left[\sum_j v_j\left(x^*(s_i,\theta_{-i});\theta_j\right)\right]. \tag{9}$$

Since x^* maximizes the sum of utilities, it must be the case that $s_i^{(k)}(\theta_i) = \theta_i$. However, if an agent does not expect his opponents to report their types truthfully, he will not reveal his true type either. We start the following section describes with the respective example.

[14] In this respect, the AGV mechanism improves over the VCG mechanism (Vickrey, Clarke, and Groves), that does not achieve budget balance (Green and Laffont 1979, Walker 1980).

[15] Recall that we assume strict concavity of $v_i(x,\theta_i)$ in x.

3 Results

Example Consider a setting with n players and a quadratic utility representation $v_i(x, \theta_i) = \theta_i x - \frac{x^2}{2}$, $i \in I$. In this setup, agent i has a bliss point at θ_i and incurs quadratic loss as the allocation departs from it. It is easy to verify that the socially efficient allocation (that maximizes the sum of utilities) is the average of individual bliss points: $x^*(\theta_1) = \frac{\sum_i \theta_i}{n}$.[16] In the appendix, we prove the following simple lemma:

Lemma 2 In the quadratic case, the optimal *Lk* strategy, $k \geq 1$, for player i is given by the following:

$$s_i^{(k)}(\theta_i) = \theta_i + \Delta \times \left(-\frac{n-1}{n}\right)^{k+1}, \tag{10}$$

where $\Delta = \int \theta dF(\theta) - \int s d\Phi(s)$ denotes the difference between the average type and the average random move of an $L0$ player.

The *Lk* strategy (10) has several interesting properties. First, the size of distortion diminishes as the level of rationality k increases. As k goes to infinity, the optimal strategies converge to truth-telling. This holds for any finite-moments distributions F and Φ. Second, if the distributions have equal means, $\int \theta dF(\theta) = \int s d\Phi(s)$, then truth-telling obtains at every level of rationality, starting from $k = 1$. Third, the absolute size of the discrepancy $\Delta \times \left(\frac{n-1}{n}\right)^{k+1}$ between the true type θ and the *L*k report $s_i^{(k)}(\theta_i)$ increases in the number of players n.

Next we study these properties in a more general setup. We maintain, however, that the efficient rule is linear in (a function of) the reported types. Formally, we make the following assumption:

A5. For all $i \in I$, $s_i \frac{\partial^2 x^*}{\partial s_i \partial s_j}(s_i, s_{-i}) = 0$.

Henceforth, we refer to A5 as the 'neutrality condition'. It implies that the marginal effect of an agent's report on the efficient allocation is not influenced by the report of another agent. The condition is satisfied whenever the efficient allocation x^* is a linear combination of types, in particular if it is the average

[16] $n = |I|$.

of types.[17] For instance, an environment with $v_i(x, \theta_i) = -(x - \theta_i)^{2p}$, for some $p \in \mathcal{N}$, $n = 2$ satisfies A5. Neutrality is a necessary condition for the proof of our main result: Proposition 2. In section 4, we discuss the case when neutrality is violated.

Observe that the first level, *L1*, is central to the analysis. As we will see next, if no distortion of truth-telling appears at *L1*, then no distortion will be observed at any subsequent level. Therefore we focus on the behavior of *L1*'s in the mechanism. Once we identify the departures from truth-telling at the first level, we study whether it dissipates at higher levels and what the implications for the AGV mechanism are.

Recall that an *L1* maximizes his expected payoff under the belief that his opponent makes a random report. The *L1* optimal strategy (best reply) in the mechanism is given by:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} \mathbb{E}_{s_{-i}^{(0)}} \left[v_i \left(x^* \left(s_i, s_{-i}^{(0)} \right); \theta_i \right) + \mathbb{E}_{\theta_{-i}} [\sum_{j \neq i} v_j \left(x^* \left(s_i, \theta_{-i} \right); \theta_j \right)] \right] \tag{11}$$

where $x^*(s_i, s_{-i})$ is the efficient social choice rule defined in Equation (3). The analysis of the optimal strategy yields the following simple result.

Proposition 1 Under assumptions A1-A3, truth-telling is optimal at all levels of rationality if the distribution of random strategies Φ and the distribution of types F coincide.

Proposition 1 establishes the equivalence between the equilibrium and *Lk* predictions of the mechanism's outcome. It implies that whether the agents stop at a finite level of reasoning or engage in equilibrium thinking is irrelevant as long as the perceived distribution of random strategy coincides with the distribution of type. Proposition 1 trivially extends to the cognitive hierarchy (CH) model, because both *Lk* and CH models define *L1* equivalently. Overall, the AGV mechanism achieves efficient implementation in four models of reasoning: *Lk* and CH with truth-telling *L0*s; *Lk* and CH with random *L0*s and $F = \Phi$.

[17] Condition A5 could also be called 'linearity'. Although A5 does not imply that the social choice function is linear in *types*, it does imply linearity in some *monotone functions* of types. Both problems are equivalent.

If distributions F and Φ do not coincide, *Lk* agents do not report truthfully. Next we show that systematic biases in reports (under- or over-reporting for *all* realizations of type) occur if F and Φ are ordered in the sense of first-order stochastic dominance. F dominates Φ means that the probability that a type exceeds a given threshold is always higher than the probability that a random move exceeds the same threshold. For instance, if Φ represents a distribution of values obtained from a prior survey, and F represents the true distribution, then a dominance relation between the distributions may arise if the survey sample is biased.

Denote the first-order stochastic dominance relation by $\succ_{FOSD}$.[18] The following proposition states in which direction a level-*1* player's report is going to be distorted.

Proposition 2 Under assumptions *A1*-*A5*, the *L1*s distort their type reports upwards if $F \succ_{FOSD} \Phi$, and downwards if $\Phi \succ_{FOSD} F$.

The proof of the proposition is given in the Appendix. We start with the observation that any n-player problem can be reduced to a problem with 2 players due to the fact that the stochastic dominance relation is preserved under monotone transformations an summation of random variables. This is the content of Lemma A in the Appendix. Then, in the framework with 2 players, we analyze the first-order condition that corresponds to the payoff-maximization problem (11) to obtain the result.

The proposition states that level-*1* players systematically (that is, for every realization of type) misreport their types, if the distributions of types and of random strategies are ordered in the sense of first-order stochastic dominance. In particular, if player *i* expects player *j* to report a higher type than *j* has on average, then *i* will report a lower type than he actually has (and vice versa), even if this induces a less preferred allocation. What is the intuition behind that? In the AGV mechanism, agent *i* gets utility from the social choice based on his and *j*'s reported preferences, plus the expected payoff of agent *j* had he told the truth to the principal. Suppose first that a high type values the size of the alternative more than a low type ('positive SMC', as in *A*4.1). If agent *i* knows that agent *j* tends to over-report his preferred allocation, then – since *i* benefits from

[18]For instance, $F \succ_{FOSD} \Phi$ reads: F dominates Φ in the sense of first-order stochastic dominance.

satisfying *j's true* preferences in expectation – he would adjust the social choice downward by under-reporting himself. If higher types prefer lower alternatives, then *j's* over-reporting makes the chosen alternative lower and i over-reports to shift it back up. In either case, the level-1 player compensates the counterpart's random behavior by misreporting their types in the opposite direction.

Recall from the example of the previous section that the distortion of reports by level-*1* players feeds into the optimal strategies of level-*2* players, level-*3* and so on, whereas the size of distortion decreases and the limiting optimal strategy is truth-telling. The following proposition states a similar result for a more general setting of an arbitrary social choice rule that satisfies neutrality.

Proposition 3 Suppose that A1-A5 hold, and $F \succ_{FOSD} \Phi$ or $\Phi \succ_{FOSD} F$. Then for all i, $\lim \mathbb{E}_{\theta_i} \left| s_i^{(k)}(\theta_i) - \theta_i \right| = 0$ and $sgn\left(s_i^{(k)}(\theta_i) - \theta_i\right) = -sgn\left(s_i^{(k-1)}(\theta_i) - \theta_i\right)$.

The expected absolute deviation of reported from true types decreases with the level of rationality. The sign of the expected deviation alternates as the level of rationality increases by one. Thus the optimal level-k strategies follow a similar pattern as the example of Section 2. If level-*2* players overstate their type in the game, then level-*3* players will understate them. Note that this is good news for the AGV mechanism: if the group of agents is a mix of, say, level-*2* and level-*3* players, then the expected chosen alternative will be closer to the one maximizing the true welfare.

4 Extension

The assumption of neutrality implies that the marginal effect of an agent's report on the allocation choice is unaffected by another agents' report. However, in certain preference environments, this assumption may be violated. For instance, if agent i of an extreme type knows that his biased report affects the mechanism's reaction to *j*'s report in such a way that the total distortion becomes even stronger, he may prefer not to misreport in the direction that Proposition 2 suggests.

This can be demonstrated by the following example. Suppose that the agents' preferences are given by $v_i = \theta_i x$, where the allocation x takes values 0 or 1 (whether or not to build an airport), and types range between -10 and 10. This

implies that, when there are two agents in the mechanism, the optimal decision is to undertake the project, $x^* = 1$, if $\theta_1 + \theta_2 > 0$ and decline, $x^* = 0$, otherwise. Suppose that F dominates Φ, and it holds for both distributions that the mass on the negative side of the support $(-10, 0)$ is very small, and the mass on the positive side $(0, 10)$ is very large (a small minority suffers from having the airport around, while a large majority benefits).

$$F(x) = \begin{cases} \varepsilon & x \in (-10, 0), \\ 1/10 - \varepsilon & x \in (0, 10), \end{cases} \qquad \Phi(x) = \begin{cases} 2\varepsilon & x \in (-10, 0), \\ 1/10 - 2\varepsilon & x \in (0, 10). \end{cases}$$

Proposition 2 says that due to the dominance relation between F and Φ $L1$ players will tend to over-report their types. Consider however an agent of type -10, who overstates his type and reports -9. This raises the probability of project implementation from 0 to $1/10 - 2\varepsilon \equiv \pi$. The expected externality equals 9.5π while i's expected utility is -10π such that his total payoff in the mechanism is negative.[19] Thus, contrary to what Proposition 2 suggests, the agent is strictly better off by reporting his type truthfully (in which case he gets the zero payoff). By over-reporting his type he will increase the probability that the project is undertaken.

The result of Proposition 2 does not apply in this example since the agents' reports are perfect substitutes when one agent's type is the negative of the other: $\theta_1 = -\theta_2$. Similar problem arises when types are complements. The formal definitions are as follows.

Agents' types are **complements**[20] with respect to the efficient allocation, if:

$$\frac{\partial^2 x^*}{\partial s_i \partial s_j}(s_i, s_j) > 0.$$

[19] Recall that we omit the budget-balancing term of the AGV transfer, since it does not affect strategy choice. The agent's payoff in the mechanism is given by $u_i = \theta_i (1 - \Phi(-s_i)) + \int_{-s_j}^{10} \theta_j dF(\theta_j)$.

[20] E.g.: $v_i(x, \theta_i) = \theta_i x - \frac{1}{x}$, $x > 0$, $\theta < 0$.

Agents' types are **substitutes**[21] with respect to the efficient allocation, if:

$$\frac{\partial^2 x^*}{\partial s_i \partial s_j}(s_i, s_j) < 0.$$

If types are substitutes, a higher report by agent i lowers the marginal effect of the opponent's report. If types are complements, the interaction is the opposite: the marginal effect of j's report increases with the report of agent i.

The following propositions state results that parallel Proposition 2 in the mechanism with two players. In this part of the analysis, we distinguish between *positive* (A4.1) and *negative* (A4.2) single crossing. Recall that, in the positive case, higher types receive higher marginal utility from allocation. In the negative case, the marginal utility diminishes with type. We separate the environments into four groups according to two criteria: first, whether the single-crossing holds as *positive* or as *negative*, and, second, whether the chosen alternative's increment due to an increase in one agent's report *increases* or *decreases* with the other agent's report (types are complements or substitutes). In these propositions, we additionally assume the monotone likelihood ratio property (*MLRP*).

Proposition 4 Under assumptions *A*1-*A*4.1, *MLRP* and complements (substitutes) environment, the agents with sufficiently low (high) types distort their reports downwards, if $\Phi \succ_{FOSD} F$, and upwards, if $F \succ_{FOSD} \Phi$.

Proposition 5 Under assumptions *A*1-*A*4.2, *MLRP* and complements (substitutes) environment, the agents with sufficiently high (low) types distort their reports downwards, if $\Phi \succ_{FOSD} F$, and upwards, if $F \succ_{FOSD} \Phi$.

Propositions 4 and 5 make four distinct claims. Let us consider, for example, the first claim: if high types tend to have high valuations (A4.1: positive single-crossing) and the efficient social choice rule is more sensitive to i's type if j's type is high (i.e., types are complements), then low-valuation players will tend to misreport their type so as to compensate the bias in the other player's report. This claim is the same as Proposition 2, except that high-valuation agents are excluded. If there is first-order stochastic dominance in distributions, in the neutral case, agent i displays compensating behavior: i systematically under- or over-reports, regardless of whether his true type is high or low. However, in a

[21] E.g.: $v_i(x, \theta_i) = \theta_i x + \frac{1}{x}$, $x < 0$, $\theta > 0$.

non-neutral case this is different. Continuing with the first claim for illustration, we observe that under its conditions the mechanism becomes more sensitive to *j's* misreporting in the range where *i's* type is high. Therefore *i*'s compensating reporting strategy has an additional indirect effect on the allocation choice (see proof in the Appendix). For this reason, both propositions 4 and 5 include only the type ranges that correspond to sufficiently weak sensitivity of the social choice rule to the other agent's report. Types in the weak-sensitivity regions display compensating behavior.

5 *Lk* as a Learning Algorithm

Observe that the *Lk* model can be thought of a model of learning. Suppose that a symmetric incomplete information game is played repeatedly by a large number of players. There is a common prior over types and types are independent. At the end of each repetition, the players observe each others' actions and types. At date 0, each player chooses a random action. At date 1, each player best-replies to the profile of actions played at date 0. At every subsequent date k, each player best-replies to his opponents' strategies at *k-1*.[22] Note that this implies that all the players play the same strategy as function of type (not the same action). Observe that the strategy played at a given date k (by all players) corresponds to an *Lk* strategy.[23] Therefore, Proposition 3 implies that such learning procedure converges to truth-telling in the AGV mechanism.

Does truth-telling convergence obtain with other learning procedures? For complete information games, Monderer and Shapley (1996) described a learning algorithm that is similar to the above interpretation of the *Lk* model. In their learning algorithm, called improvement path, one player improves his payoff at a given date k, while the rest play as in *k-1*. In the appendix, we extend the improvement path algorithm to games of incomplete information and show that the game induced by the AGV mechanism with quadratic utility $v_i(x, \theta_i) = \theta_i x - \frac{x^2}{2}$ is a potential game for any type profile θ. Applying the result of Monderer and Shapley we conclude that in the quadratic case the improvement path leads to

[22] If the number of players is large and the game is symmetric, strategies can be inferred from the observed actions and types.

[23] Note that the learning interpretation of the cognitive hierarchy model is closer to fictitious play, since it takes into account the weighted average of the whole past, and not just the last period.

truth-telling in the AGV mechanism. This, however, is not true in general: even in neutral concave environments that are not quadratic the AGV does *not* induce a potential game for all type profiles θ.[24] This latter finding relates to Sandholm (2005) who designs an indirect mechanism with the potential game property and shows that convergence to efficiency obtains in a very large class of learning dynamics. Lifting the concavity assumption made in this paper, Mathevet (2010) designs supermodular mechanisms with good learning properties. When a game is potential or supermodular, then a large class of learning algorithms converge to equilibrium.

6 Conclusion

The idea of relaxing the pervasive common knowledge assumption, often referred to as the Wilson doctrine, has motivated recent research in mechanism design. Significant progress was made in studying implementation in frameworks approaching the universal type space, where higher-order beliefs are virtually unrestricted.[25] Kets (2012) extends the notion of type space further to allow finite depths of reasoning, as in the level-k model. The next natural step for mechanism design is to accommodate the extended notion and search for mechanisms that are robust with respect to changes not only in the structure of beliefs, but also in the depth of reasoning (as mentioned in the discussion, learning to play the mechanism is a related issue). This paper first studies one of the most influential of existing mechanisms, d'Aspremont and Gerard-Varet (1979), in the *Lk* environment.

The AGV mechanism implements the efficient choice rule under the common prior and common knowledge assumptions. It is conceptually similar to the Vickrey-Clarke-Groves (VCG) mechanism that taxes the agents with the amount of negative externality their preference report exerts on the welfare of other agents. The VCG mechanism implements the efficient social choice rule in dominant strategies, and hence is independent of the beliefs.[26] On the downside, the

[24] Observe that this implies that our result of convergence to truthtelling in the *Lk* model is *not* due to the potential game property.

[25] This literature stems from Bergemann and Morris (2005).

[26] Dominant-strategy implementation guarantees that the VCG mechanism achieves truthful revelation and efficiency, for any $k > 0$ in the Lk model.

VCG mechanism fails to satisfy the overall budget constraint. The expected externality mechanism has the advantage of being exactly budget balanced, but it comes at the cost of achieving Bayesian, as opposed to dominant-strategy implementation. In the light of the L*k* model, this is not entirely innocuous.

We show that if there is a systematic difference in the perceptions of random-L0 moves and the true types the agents will distort their types at the first level and, by extension, also at the higher levels of rationality. We observe compensating behavior of finite-level players in an AGV mechanism, that is, distorting one's report in the opposite direction to the anticipated bias of the opponents. This is due to the fact that the AGV mechanism rewards for the expected externality, where the expectation is measured with respect to the true types. A simple implication of this result is that the AGV mechanism could use the distribution of random moves, as opposed to types, to achieve truthtelling among *Lk* agents.

Nevertheless our results, put together, vindicate the AGV mechanism in convex environments. First, in the truthful-*L0* specification there is no distortion of truthtelling and efficiency. Second, if there is distortion of truth-telling, its sign alternates and its absolute value decreases with k. Therefore, in mixed groups of agents with various levels k the biases cancel out and the mechanism's outcome is close to efficiency. This also implies that starting from *L2* in the cognitive hierarchy model best replies are located within a smaller neighborhood of truth-telling. Third, our convergence result suggest that, in repeated interactions where the agents can observe others' strategies, equilibrium becomes an increasingly better approximation and the expected externality mechanism achieves efficiency.

A Appendix

Lemma 2

Statement $s_i^{(k)}(\theta_i) = \theta_i + \Delta \times \left(-\frac{n-1}{n}\right)^{k+1}$, $k \geq 1$, where $\Delta = \int \theta dF(\theta) - \int s d\Phi(s)$.

Proof We proceed by induction. Suppose that for $k-1$ it holds that:

$$s^{(k-1)}(\theta_j) = \theta_i + \left(-\frac{n-1}{n}\right)^k \Delta \tag{12}$$

Level-k optimal strategy is best reply to the profile of strategies $s^{(k-1)}(\theta_j)$, where the expectation is taken with respect to the opponents' types θ_{-i}.

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} \mathbb{E}_{\theta_{-i}} \left[\theta_i \left(\frac{s_i + \sum_{j \neq i} s^{(k-1)}(\theta_j)}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} s^{(k-1)}(\theta_j)}{n} \right)^2 + \right.$$

$$\left. + \sum_{j \neq i} \mathbb{E}_{\theta_{-i}} \left[\theta_{-i} \left(\frac{s_i + \sum_{j \neq i} \theta_{-i}}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} \theta_{-i}}{n} \right)^2 \right] \right] = \theta_i + \frac{n-1}{n} \left(\mathbb{E}\theta_j - \mathbb{E}s^{(k-1)}(\theta_j) \right)$$

$$= \theta_i + \frac{n-1}{n} \left(\mathbb{E}\theta_j - \mathbb{E} \left[\theta_j + \left(-\frac{n-1}{n} \right)^k \Delta \right] \right) = \theta_i + \left(-\frac{n-1}{n} \right)^{k+1} \Delta$$

Thus, if (12) holds on level $k-1$ it also holds on level k. Level-1 strategy is best reply to the profile of random moves:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} \mathbb{E}_{s_{-i}^{(0)}} \left[\theta_i \left(\frac{s_i + \sum_{j \neq i} s_j^{(0)}}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} s_j^{(0)}}{n} \right)^2 + \right.$$

$$\left. + \sum_{j \neq i} \mathbb{E}_{\theta_{-i}} \left[\theta_j \left(\frac{s_i + \sum_{j \neq i} \theta_j}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} \theta_j}{n} \right)^2 \right] \right] = \theta_i - \frac{n-1}{n} \Delta,$$

Thus for *L1* the induction formula (12) applies. $(Lemma\,2)$■

Proposition 1

Statement Under assumptions A1-A3, if $F \equiv \Phi$ then $s_i^{(k)}(\theta_i) = \theta_i$ for all k, $i \in I$.

Proof The first-order condition (henceforth f.o.c.) for the maximization problem (11) is the following:

$$\mathbb{E}_{s_{-i}^{(0)}} \left[\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); \theta_i) \frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)}) \right] + \sum_{j \neq i} \mathbb{E}_{\theta_{-i}} \left[\frac{\partial v_j}{\partial x}(x^*(s_i, \theta_{-i}); \theta_j) \frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i}) \right] = 0 \tag{13}$$

Given that $x^*(s_i, s_{-i})$ is the efficient choice rule, it must hold that

$$\sum_{j \neq i} \frac{\partial v_j}{\partial x}(x^*(s_i, \theta_{-i}); \theta_j) + \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i) = 0.$$

Then the second term of (13) can be rewritten, such that the f.o.c. becomes:[27]

$$\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(0)}\right); \theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(0)}\right)\right] - \mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}\left(x\left(s_i, \theta_{-i}\right); s_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, \theta_{-i}\right)\right] = 0 \tag{14}$$

Therefore, if $F(t) = \Phi(t)$ (i.e. $s_{-i}^{(0)}$ and θ_{-i} is the same random variable), then $s_i = \theta_i$ satisfies the first order condition (14) and thus $s_i^{(1)}(\theta_i) = \theta_i$. $(P1)\blacksquare$

Lemma A

Statement Suppose A1-A5 hold. Consider an *L1* problem P_n with n players and $F \prec_{FOSD} \Phi$ ($\Phi \prec_{FOSD} F$). There exists an *L1* problem P_2 with 2 players and a pair of distribution functions F^Σ, Φ^Σ satisfying $F^\Sigma \prec_{FOSD} \Phi^\Sigma$ ($\Phi^\Sigma \prec F^\Sigma$) such that the solution to P_2 is also a solution to P_n.

Proof First, we observe that $\frac{\partial^2 x^*}{\partial s_i \partial s_j} = 0$ (A5) implies that $x^*(s_1, ..s_n) = \sum_i \lambda_i s_i$ for some scalars λ_i, $\lambda_i > 0$.[28] Condition (14) can be rewritten as follows:

$$\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(\sum_{j \neq i}\lambda_j s_j^{(0)} + \lambda_i s_i; \theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(0)}\right)\right]$$

$$= \mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}\left(\sum_{j \neq i}\lambda_j \theta_j + \lambda_i s_i; s_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, \theta_{-i}\right)\right].$$

[27] The second order condition (s.o.c.) $\mathbb{E}_{s_{-i}^{(0)}}[\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, s_{-i}^{(0)}); \theta_i)[\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})]^2 + \frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial^2 x^*}{\partial s_i^2}(s_i, s_{-i}^{(0)})] -$
$-\mathbb{E}_{\theta_{-i}}[\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \theta_{-i}); s_i)[\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})]^2 + \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i)\frac{\partial^2 x^*}{\partial s_i^2}(s_i, \theta_{-i}) + \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \theta_{-i}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})]|_{\substack{s_i = \theta_i \\ F(\cdot) = \Phi(\cdot)}} =$
$= -\mathbb{E}_{\theta_{-i}}[\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \theta_{-i}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})] < 0$ (see Lemma C).

[28] This assumes that types are relabeled: if $v_i(x, \theta_i) \equiv \tilde{v}_i(x, h(\theta_i))$, then we consider $\tilde{v}_i\left(x, \tilde{\theta}_i\right)$ with type $\tilde{\theta}_i = h(\theta_i)$. If A4.2 holds, ("negative" SMC), let $\tilde{\theta}_i = -h(\theta_i)$.

s_i that satisfies this condition is a solution to P_n. From Theorem 1.A.3 in Shaked and Shanthikumar (2007): if distribution Φ of $s_j^{(0)}$ dominates distribution F of θ_j, then distribution Φ^Σ of $s_\Sigma^{(0)} \equiv \sum_{j\neq i}\lambda_j s_j^{(0)}$ dominates distribution F^Σ of $\theta_\Sigma \equiv \sum_{j\neq i}\lambda_j\theta_j$, and vice versa. $s_\Sigma^{(0)}$ and θ_Σ correspond to the random action and type of a fictitious second player in P_2. In this problem P_2 the first order condition writes as follows:

$$\mathbb{E}_{s_\Sigma^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(s_\Sigma^{(0)}+\lambda_i s_i;\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_\Sigma^{(0)}\right)\right]$$

$$=\mathbb{E}_{\theta_\Sigma}\left[\frac{\partial v_i}{\partial x}\left(\theta_\Sigma+\lambda_i s_i;s_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,\theta_\Sigma\right)\right].$$

It is then clear that the solutions to problems P_n and P_2 coincide.

$(Lemma\ A)\blacksquare$

Lemma B.

Statement. The *L1* strategy in the AGV mechanism is given by ($n=2$):

$$s_i^{(1)}(\theta_i)=\theta_i+\frac{\int\left(F(t)-\Phi(t)\right)d\frac{\partial v_i}{\partial x}(x^*(s_i^{(1)}(\theta_i),t);s_i^{(1)}(\theta_i))\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i),t)}{\int\frac{\partial^2 v_i}{\partial x\partial\theta_i}(x^*(s_i^{(1)}(\theta_i),s_{-i}^{(0)});\widehat{\theta}_i)\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i),s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})} \tag{15}$$

Proof. Rewrite (14) as follows:

$$0=\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}(x^*(s_i,s_{-i}^{(0)});\theta_i)\frac{\partial x^*}{\partial s_i}(s_i,s_{-i}^{(0)})\right]-\mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})\right]$$

$$=\int\frac{\partial v_i}{\partial x}(x^*(s_i,s_{-i}^{(0)});\theta_i)\frac{\partial x^*}{\partial s_i}(s_i,s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})-\int\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})dF(\theta_{-i}) \tag{16}$$

Integrate the second term of Equation (16) by parts:

$$\int\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})dF(\theta_{-i})=$$

$$=\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})F(\theta_{-i})\Big|_\Theta-\int F(\theta_{-i})d\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})$$

Modify the first term of Equation (16) by taking Taylor expansion under the integral:

$$\int \frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) =$$

$$= \int [\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); s_i) + \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)(\theta_i - s_i)]\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})$$

where $\widehat{\theta}_i$ is between s_i and θ_i,

$$= \int \frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) +$$

$$+ \int \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)(\theta_i - s_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) =$$

and integrate by parts:

$$= \frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})\Phi(s_{-i}^{(0)})\Big|_\Theta - \int \Phi(t) d\frac{\partial v_i}{\partial x}(x^*(s_i, t); s_i)\frac{\partial x^*}{\partial s_i}(s_i, t) +$$

$$+ \int \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)(\theta_i - s_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})$$

Observe that due to the equal support of the two distribution functions F and Φ:

$$\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})\Phi(s_{-i}^{(0)})\Big|_\Theta = \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})F(\theta_{-i})\Big|_\Theta$$

Thus, the f.o.c. becomes:

$$\int (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i, t); s_i)\frac{\partial x^*}{\partial s_i}(s_i, t) +$$

$$+(\theta_i - s_i)\int \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) = 0$$

We can rewrite the solution as follows:

$$s_i^{(1)}(\theta_i) - \theta_i \equiv \frac{\int (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i^{(1)}(\theta_i), t); s_i^{(1)}(\theta_i))\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i), t)}{\int \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i^{(1)}(\theta_i), s_{-i}^{(0)}); \widehat{\theta}_i)\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})} \qquad (17)$$

If $F(t) - \Phi(t) \equiv 0$, then $s_i^{(1)}(\theta_i) = \theta_i$, hence the lemma. $(Lemma\ B)\blacksquare$

Lemma C

Statement The Spence-Mirrlees condition ($A4$) implies the following, for all $\theta_i, \widehat{\theta}_i, s_{-i}^{(0)}$:

$$\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i^{(1)}(\theta_i), s_{-i}^{(0)}); \widehat{\theta}_i)\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)}) > 0.$$

Proof The efficiency of the social choice rule x^* implies that for all t_i, t_{-i}:

$$\frac{\partial v_i}{\partial x}(x^*(t_i, t_{-i}), t_i) + \frac{\partial v_{-i}}{\partial x}(x^*(t_i, t_{-i}), t_{-i}) \equiv 0$$

Differentiate with respect to θ_i:

$$\frac{\partial x^*}{\partial s_i}(t_i, t_{-i})[\frac{\partial^2 v_i}{\partial x^2}(x^*(t_i, t_{-i}), t_i) + \frac{\partial^2 v_{-i}}{\partial x^2}(x^*(t_i, t_{-i}), t_{-i}] + \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(t_i, t_{-i}), t_i) = 0$$

From the s.o.c. of the same problem,

$$\frac{\partial^2 v_i}{\partial x^2}\,(x^*(t_i, t_{-i}), t_i) + \frac{\partial^2 v_{-i}}{\partial x^2}\,(x^*(t_i, t_{-i}), t_{-i}) < 0$$

Thus, $\text{sgn}(\frac{\partial x^*}{\partial s_i}(t_i, t_{-i})) = \text{sgn}(\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(t_i, t_{-i}), t_i)$. Substitute t_i by $s_i^{(1)}(\theta_i)$, t_{-i} by $s_{-i}^{(0)}$ and obtain:

$$\text{sgn}(\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)})) = \text{sgn}(\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i^{(1)}(\theta_i), s_{-i}^{(0)}), s_i^{(1)}(\theta_i))).$$

Given $A4$ (i.e., sign of $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x, \theta_i)$ is the same for all (x, θ_i)) the result is proven. $(Lemma\, C)$■

Proposition 2.

Statement Suppose A1-A5 hold. If $F \succ_{FOSD} \Phi$ then $s_i^{(1)}(\theta_i) > \theta_i$, and if $\Phi \succ_{FOSD} F$ then $s_i^{(1)}(\theta_i) < \theta_i$.

Proof From Lemma B, the first-order condition for the $L1$ maximization problem when $n = 2$ is given by Equation (17). Lemma C (p. 23) shows that the denominator of the expression is positive. Let us transform the nominator as follows:

$$\int (F(t)-\Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) =$$

$$= \int (F(t)-\Phi(t))\,[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-(1)}\underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+(2)}+$$

$$+\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)}_{=0(3)}]dt$$

The signs marked above are determined by the following.

(1) $\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i) < 0$ by the concavity of preferences;

(2) By Lemma C (p. 23), $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*;\theta_i)\frac{\partial x^*}{\partial s_i} > 0$ for all i, θ_i, s_i, s_{-i}; by A4, the signs of $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*;\theta_i)$ and $\frac{\partial^2 v_{-i}}{\partial x \partial \theta_{-i}}(x^*;\theta_{-i})$ are invariant for all θ_i, s_i, s_{-i};

(3) $\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t) = 0$ by neutrality.

Therefore, the term

$$[\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_{-i}}(s_i,t)\frac{\partial x^*}{\partial s_i}(s_i,t) + \frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)]$$

is negative. Given that $\Phi \succ F$ implies $F(t)-\Phi(t) > 0$ for all t and $\Phi \prec F$ implies $F(t)-\Phi(t) < 0$ Proposition 2 follows immediately. $(P2)\blacksquare$

Proposition 3.

Statement Suppose that A1-A5 hold, and $F \succ_{FOSD} \Phi$ or $\Phi \succ_{FOSD} F$. Then for all i, $\lim \mathbb{E}_{\theta_i}\left[\left|s_i^{(k)}(\theta_i)-\theta_i\right|\right] = 0$ and $sgn\left(s_i^{(k)}(\theta_i)-\theta_i\right) = -sgn\left(s_i^{(k-1)}(\theta_i)-\theta_i\right)$.

Proof Recall that by definition:

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} \mathbb{E}_{\theta_{-i}}[v_i(x^*(s_i, s_{-i}^{(k-1)}(\theta_{-i})),\theta_i) + v_{-i}(x^*(s_i,\theta_{-i});\theta_{-i})]$$

The first-order condition for level-k strategy $s_i^{(k)}(\theta_i)$ is as follows ($s_i^{(k)}(\theta_i) = s_i$):[29]

$$\begin{aligned}
0 &= \mathbb{E}_{\theta_{-i}}[\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) + \\
&\quad + \frac{\partial v_{-i}}{\partial x}(x^*(s_i, \theta_{-i}); \theta_{-i})\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})] \\
&= \mathbb{E}_{\theta_{-i}}\left[\left[\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i}))\right] + \right. \\
&\quad \left. - \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})\right] \\
&\overset{(*)}{=} \mathbb{E}_{\theta_{-i}}\left[(\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) - \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i))\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) + \right. \\
&\quad \left. + \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i)\underbrace{(\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) - \frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i}))}_{=0}\right].
\end{aligned}$$

$\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) - \frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i}) = 0$ since by neutrality assumption $\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i, t) = 0$ and $x^*(\cdot, \cdot)$ is continuously differentiable.

Apply the Taylor expansion to the first term:

$$\begin{aligned}
0 = \mathbb{E}_{\theta_{-i}}&\left[\left(\frac{\partial v_i}{\partial x}(x^*(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) - \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i)\right)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i}))\right] \\
&= \mathbb{E}_{\theta_{-i}}\left[\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i}) + \right. \\
&\quad \left. + \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})(\theta_i - s_i)\right]\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i}))
\end{aligned}$$

where $\widehat{\theta_i} \in [\min(\theta_i, s_i); \max(\theta_i, s_i)]$, and $\widehat{s_{-i}} \in \left[\min(s_{-i}^{(k-1)}(\theta_{-i}), \theta_{-i}); \max(s_{-i}^{(k-1)}(\theta_{-i}), \theta_{-i})\right]$

Since $\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) \neq 0$ we get:

$$s_i - \theta_i = \mathbb{E}_{\theta_{-i}}\left[\underbrace{\frac{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}}_{<0}(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})\right],$$

Recall that $s_i = s_i^{(k)}(\theta_i)$; the distortion of type changes sign as k increases by 1.

[29] To perform transition $(*)$ we add and subtract $\frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i}))$.

Remark Recall from Proposition 2 that either $s_i^{(1)}(\theta_i) \geq \theta_i \ \forall \theta_i$, or $s_i^{(1)}(\theta_i) \leq \theta_i \ \forall \theta_i$. By induction, the equation above implies that the same is true for all levels k: either $s_i^{(k)}(\theta_i) \geq \theta_i \ \forall \theta_i$, or $s_i^{(k)}(\theta_i) \leq \theta_i \ \forall \theta_i$.

Moreover, from the proof of Lemma C we know that

$$\frac{-\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}}) - \frac{\partial^2 v_{-i}}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_{-i})\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); s_i)} = 1,$$

thus $\frac{-\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_i)\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); s_i)} < 1$.[30]

For $\hat{\theta}_i$ we have, by continuity,

$$\frac{-\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \hat{\theta}_i)\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \hat{\theta}_i)} < 1$$

as well. Take the expectation of both sides:

$$\mathbb{E}_{\theta_i}\left[s_i^{(k)}(\theta_i) - \theta_i\right] = \mathbb{E}_{\theta_i}\mathbb{E}_{\theta_{-i}}\left[\frac{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})\right]$$

as types are independent and the distributions of types coincide,

$$\mathbb{E}_{\theta_i}\left[s_i^{(k)}(\theta_i) - \theta_i\right] = \mathbb{E}_{\theta_{-i}}\left[(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})\mathbb{E}_{\theta_i}\frac{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}\right]$$

$$\mathbb{E}_{\theta_i}\left|s_i^{(k)}(\theta_i) - \theta_i\right| < \mathbb{E}_{\theta_i}\left|s_i^{(k-1)}(\theta_i) - \theta_i\right| \tag{18}$$

Consider the sequence $\left\{\mathbb{E}_{\theta_i}\left|s_i^{(k)}(\theta_i) - \theta_i\right|\right\}_k$. Since $\mathbb{E}_{\theta_i}\left|s_i^{(k)}(\theta_i) - \theta_i\right| \geq 0$, inequality 18 implies that the sequence converges. The proof is by contradiction. Let $\overline{L}$ denote the limit of the sequence, and suppose $s_i^{limsup}(\cdot) > s_i^{liminf}(\cdot)$ are such that $\mathbb{E}_{\theta_i}\left(s_i^{limsup}(\theta_i) - \theta_i\right) = -\mathbb{E}_{\theta_i}\left(s_i^{liminf}(\theta_i) - \theta_i\right) = \overline{L}$ (take note of our remark on page 26). By the continuity of the best reply correspondence, strategy $s_i^{limsup}(\theta_i)$ is best reply to $s_i^{liminf}(\theta_i)$ and vice versa. Therefore, inequality 18 should apply to these strategies as well. But this generates a contradiction – thus $s_i^{limsup}(\theta_i) = s_i^{liminf}(\theta_i) = \theta_i$ (and $\overline{L} = 0$).

[30] $\frac{-\frac{\partial^2 v_{-i}}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_{-i})\frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); s_i)} \in]0, 1[$.

This concludes the proof of Proposition 3. $(P3)\blacksquare$

Proposition 4. Let us separate the statements of Proposition 4 and refer to them as Proposition 4a and Proposition 4b respectively. Bold face is used to emphasize the differences in the two statements:

Proposition 4a: Under A1-A4.1, MLRP and **complements** environment, $\exists t^*$ such that for all $\theta_i < t^*$ if $\Phi \succ F$ then $s_i^{(1)}(\theta_i) < \theta_i$, and if $F \succ \Phi$ then $s_i^{(1)}(\theta_i) > \theta_i$.

Proposition 4b: Under A1-A4.1, MLRP and **substitutes** environment, $\exists t^*$ such that for all $\theta_i > t^*$ if $\Phi \succ F$ then $s_i^{(1)}(\theta_i) < \theta_i$, and if $F \succ \Phi$ then $s_i^{(1)}(\theta_i) > \theta_i$.

Proof Given the non-neutrality, $\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)$, we need to decompose the denominator of Equation 17. Start with the case of Proposition 4a:
$\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x,\theta_i) > 0$, $\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t) \geq 0$. The nominator:

$$\int_{\underline{t}}^{+\infty} (F(t)-\Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) =$$

$$\underbrace{\int_{s_i}^{+\infty} (F(t)-\Phi(t)) \left[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-(1)} \underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)}_{+(2)} \underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+(2)} + \underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{-(3)} \underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)}_{+(4)} \right] dt}_{\text{"first term"}} +$$

$$\underbrace{\int_{\underline{t}}^{s_i} (F(t)-\Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)}_{\text{"second term"}}$$

It is convenient to separate the integral into two parts since $\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)$ decreases in t[31] and $\frac{\partial v_i}{\partial x}(x^*(s_i,s_i);s_i) = 0$. Consider *the first term* in brackets:

(1) $\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i) < 0$ by the concavity assumption

(2) $\frac{\partial x^*}{\partial s_{-i}}(s_i,t) > 0$, $\frac{\partial x^*}{\partial s_i}(s_i,t) > 0$ from A4.1 and Lemma C

(3) $\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i) < 0$ for $t \leq s_i$

(4) $\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t) > 0$ by the complementarity.

[31] $\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial t}(s_i,t) < 0$.

Thus we obtain that

$$\left[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-}\underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)}_{+}\underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+}+\underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{-}\underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)}_{+}\right]$$

is negative. *The second term* can be rewritten as follows:

$\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) =$

$$= \underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{=0}\frac{\partial x^*}{\partial s_i}(s_i,s_i)\,(F(s_i)-\Phi(s_i)) + \frac{\partial v_i}{\partial x}(x^*(s_i,\underline{t});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\underline{t})\underbrace{(F(\underline{t})-\Phi(\underline{t}))}_{=0}$$
$$-\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)d\,(F(t)-\Phi(t))$$

$$= -\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)\,(f(t)-\varphi(t))\,dt,$$

where

$$\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\geq 0,$$

for $t \leq s_i$

$$\frac{\partial x^*}{\partial s_i}(s_i,t) > 0.$$

First, suppose $\Phi \succ_{FOSD} F$: $F(t)-\Phi(t) > 0\ \forall t \Rightarrow$ the first term is negative. If $f(s_i)-\varphi(s_i) > 0$, then the second term is negative, too: By the MLRP assumption, $\frac{f(t)}{\varphi(t)}$ decreases in t; thus, there exists a t^* such that $f(t^*)-\varphi(t^*)=0$. This implies that, for θ_i such that $s_i^{(1)}(\theta_i) \leq t^*$, the result is established: the *L1*s with sufficiently low types distort there reports downwards.

Now suppose that $F \succ_{FOSD} \Phi$. Then, the first term is positive. By MLRP, $\frac{\varphi(t)}{f(t)}$ decreases in t and by the same reasoning for θ_i low enough the second term is positive, too, hence type reports are distorted upwards.

Proposition 4a is now proven. $(P4a)\blacksquare$

To prove Proposition 4b ($\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t) \leq 0$), we change the decomposition of the nominator as follows:

$\int_{\underline{t}}^{+\infty} (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) =$

$$\int_{\underline{t}}^{s_i} (F(t) - \Phi(t)) [\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-}\underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)}_{+}\underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+}$$
$$+\underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{+}\underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)}_{-}]dt$$
$$+\int_{s_i}^{+\infty} (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) \tag{19}$$

Given that $\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)$ decreases in t, we have that for $t \leq s_i$, $\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i) \geq 0$ and thus the term in brackets is negative. Integrating the second term by part, we obtain:

$$-\int_{s_i}^{+\infty} \underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{-}\underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+} (f(t) - \varphi(t))\, dt.$$

Similarly to the argument in 4a, we identify the condition under which both parts of the nominator have the same sign. Given the decomposition (19), we can see that for this to hold s_i has to be sufficiently *high* (or θ_i such that $s_i^{(1)}(\theta_i) \geq t^*$). Proposition 4b proven. $(P4b)\blacksquare$

$(P4)\blacksquare$

Proof of Proposition 5. The statement and proof are symmetric to Proposition 4.

***Lk* as a Learning Algorithm** Recall that in AGV, the payoff of player i with type θ_i is the following:

$$\begin{aligned} u_i(s,\theta_i) &= v_i(x^*(s),\theta_i) + \mathbb{E}_{\theta_{-i}} \sum_{j\neq i} v_j(x^*(s_i,\theta_{-i});\theta_j) \\ &\quad -\frac{1}{n-1}\sum_{l\neq i}\mathbb{E}_{\theta_{-l}}\sum_{j\neq l} v_j(x^*(s_l,\theta_{-l});\theta_j). \end{aligned} \tag{20}$$

The marginal payoff as function of own type report s_i equals:

$$\frac{\partial u_i}{\partial s_i}(s,\theta_i) = \frac{\partial v_i}{\partial x}(x^*(s);\theta_i)\frac{\partial x^*}{\partial s_i}(s) + \sum_{j\neq i}\mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_j}{\partial x}(x^*(s_i,\theta_{-i});\theta_j)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})\right]. \quad (21)$$

Taking the derivative of (21) with respect to s_j we obtain:

$$\begin{aligned}\frac{\partial^2 u_i}{\partial s_i \partial s_j}(s,\theta_i) = {} & \frac{\partial^2 v_i}{\partial x^2}(x^*(s);\theta_i)\frac{\partial x^*}{\partial s_j}(s)\frac{\partial x^*}{\partial s_i}(s) + \frac{\partial v_i}{\partial x}(x^*(s);\theta_i)\frac{\partial^2 x^*}{\partial s_i \partial s_j}(s) \\ & + \sum_{j\neq i}\mathbb{E}_{\theta_{-i}}\frac{\partial^2 v_j}{\partial x^2}(x^*(s_i,\theta_{-i});\theta_j)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})\frac{\partial x^*}{\partial s_j}(s_i,\theta_{-i}) \\ & + \sum_{j\neq i}\mathbb{E}_{\theta_{-i}}\frac{\partial v_j}{\partial x}(x^*(s_i,\theta_{-i});\theta_j)\frac{\partial^2 x^*}{\partial s_i \partial s_j}(s_i,\theta_{-i}). \quad (22)\end{aligned}$$

From Monderer, Shapley (1996) we know that a complete information game with payoffs $(u_i)_{i\in I}$ is an (exact) potential game if and only if $\frac{\partial^2 u_i}{\partial s_i \partial s_j} = \frac{\partial^2 u_j}{\partial s_i \partial s_j}$ for every i and j (Theorem 4.5). We assume that the utility functions $v_i(\cdot)$ and type distributions are the same for all i, and denote $\Gamma(\theta)$ the game with payoffs given by (20).

It follows immediately from equation (22) that game $\Gamma\left(\overline{\theta}\right)$ where $\overline{\theta_i} = \overline{\theta_j}$ for all i and j is then a potential game. However, naturally, we are not interested in the particular realization $\overline{\theta}$ of type profile, but rather in all possible type profiles – that is, in the family of games $\{\Gamma(\theta) : \theta \in \Theta^n\}$. Let us call an incomplete information game an *ex post potential game* if it is a potential game for all realizations of types. Formally, $\Gamma = \left\{I, (S_i, \Theta_i, F_i, u_i)_{i\in I}\right\}$ is an ex post potential game if for all $\theta_i \in supp(F_i)$, game $\Gamma(\theta) = \left\{I, (S_i, u_i(\cdot,\theta_i))_{i\in I}\right\}$ has a potential.

If payoffs in game $\Gamma(\theta)$ are given by (20) where $v_i(x,\theta_i) = \theta_i x - \frac{x^2}{2}$ (as in the example of Section 2) then, by Theorem 4.5, $\Gamma(\theta)$ is a potential game, since $\frac{\partial^2 u_i}{\partial s_i \partial s_j}(s,\theta_i) = \frac{1}{n}$ for all i, j. Thus, under such preferences, the AGV mechanism is an ex post potential game. Interestingly, the linearity of the choice rule x^* is *not* sufficient for AGV to be an ex post potential game. In the case of linearity and symmetry, equation (22) becomes:

$$\frac{\partial^2 u_i}{\partial s_i \partial s_j}(s,\theta_i) = \frac{\partial^2 v}{\partial x^2}(x^*(s);\theta_i)\frac{1}{n^2} + \mathbb{E}_{\tilde{\theta}_{-i}}\frac{\partial^2 v}{\partial x^2}\left(x^*(s_i,\tilde{\theta}_{-i});\tilde{\theta}_j\right)\frac{1}{n} \quad (23)$$

Clearly, if $\frac{\partial^2 v}{\partial x^2}\left(x^*(s);\theta_i\right)$ varies with θ_i then so does $\frac{\partial^2 u_i}{\partial s_i \partial s_j}\left(s,\theta_i\right)$; therefore, the conditions of Theorem 4.5 are not satisfied.

References

[1] Susan Athey and Ilya Segal. An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485, 2013.

[2] Pierpaolo Battigalli and Marciano Siniscalchi. Rationalizable bidding in first-price auctions* 1. *Games and Economic Behavior*, 45(1):38–72, 2003.

[3] Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, 73(6):pp. 1771–1813, 2005.

[4] Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, 2010.

[5] Colin F. Camerer and Teck-Hua Ho. Chapter 10 - behavioral game theory experiments and modeling. volume 4 of *Handbook of Game Theory with Economic Applications*, pages 517 – 573. Elsevier, 2015.

[6] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

[7] Kim-Sau Chung and J. C. Ely. Foundations of dominant-strategy mechanisms. *The Review of Economic Studies*, 74(2):pp. 447–476, 2007.

[8] Miguel Costa-Gomes, Vincent P Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001.

[9] Miguel A Costa-Gomes and Vincent P Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, 2006.

[10] Vincent Crawford, Tamar Kugler, Zvika Neeman, and Ady Pauzner. Behaviorally Optimal Auction Design: Examples and Observations. *Journal of the European Economic Association*, 7(2-3):377–387, 2009.

[11] Vincent P Crawford and Nagore Iriberri. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770, 2007.

[12] Claude d'Aspremont and Louis-Andre Gerard-Varet. Incentives and incomplete information. *Journal of Public Economics*, 11(1):25 – 45, 1979.

[13] Eddie Dekel, Drew Fudenberg, and David K Levine. Learning to play bayesian games. *Games and Economic Behavior*, 46(2):282–303, 2004.

[14] Jerry R Green and Jean-Jacques Laffont. Incentives in public decision making. 1979.

[15] Melvin J Guyer and Anatol Rapoport. $2\times$ 2 games played once. *Journal of Conflict Resolution*, 16(3):409–431, 1972.

[16] Josef Hofbauer and William H Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

[17] Willemien Kets. Bounded reasoning and higher-order uncertainty. *Available at SSRN 2116626*, 2012.

[18] Dorothea Kuebler and Georg Weizsaecker. Limited depth of reasoning and failure of cascade formation in the laboratory. *The Review of Economic Studies*, 71(2):pp. 425–441, 2004.

[19] Laurent Mathevet. Supermodular mechanism design. *Theoretical Economics*, 5(3):403–443, 2010.

[20] James A Mirrlees. An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2):175–208, 1971.

[21] Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):pp. 1313–1326, 1995.

[22] William H Sandholm. Negative externalities and evolutionary implementation. *The Review of Economic Studies*, 72(3):885–915, 2005.

[23] William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

[24] Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer, 2007.

[25] Dale O. Stahl and Paul W. Wilson. Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25(3):309 – 327, 1994.

[26] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[27] Mark Walker. On the nonexistence of a dominant strategy mechanism for making optimal public decisions. *Econometrica: Journal of the Econometric Society*, pages 1521–1540, 1980.

[28] Robert Wilson. Game-theoretic approaches to trading processes. In *Advances in Economic Theory*.