The expected externality mechanism in a level-k environment

Olga Gorelkina

International Journal of Game Theory

ISSN 0020-7276

Int J Game Theory
DOI 10.1007/s00182-017-0579-5

Int J Game Theory
DOI 10.1007/s00182-017-0579-5

ORIGINAL PAPER

The expected externality mechanism in a level-*k* environment

Olga Gorelkina[1]

Accepted: 4 May 2017

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 and Gerard-Varet, J Public Econ 11:25–45, 1979) in a finite-depth environment described by the Level-*k* model. While the original mechanism may fail to implement the efficient rule in this environment, it can be adjusted to restore efficiency.

Keywords Expected externality · Externality mechanisms · Level-*k* · Bounded rationality

JEL Classification C72 · D71 · D82

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

I am grateful to Vincent Crawford, Francoise Forges, Alia Gizatulina, Ioanna Grypari, Martin Hellwig, Rida Laraki, Thomas Mariotti, David Martimort, Benny Moldovanu, Thomas Rieck, Nicolas Roux, as well as to two anonymous referees and the Associate Editor for their helpful comments.

✉ Olga Gorelkina
olga@liv.ac.uk

[1] University of Liverpool Management School, Chatham Street, Liverpool L69 7ZH, UK

Published online: 17 May 2017

the allocation that maximizes total welfare. The major 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 Gérard–Varet (1979) construct an ingenious mechanism that aligns the agents' individual incentives with total welfare maximization. In a Bayes–Nash equilibrium, the agents reveal their types to the principal and thus efficiency can be 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 (1) whether the participants' behavioral response to the mechanism coincides with the Bayes–Nash prediction and, if it does not, (2) 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 may be negative. As to the second question, little is known as to the loss of efficiency if the participants do not play equilibrium. 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 the behavioral setting follows a large body of evidence from experimental games. Recent surveys by Crawford et al. (2009) 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 et al. 2001; Costa-Gomes and Crawford 2006) and the *cognitive hierarchy* model (CH; Camerer et al. 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.[2]

Lk is a model of reasoning prior to a game, where the agent maximizes his payoff against a non-equilibrium belief about other agents' strategies. The belief is constructed in the following iterative process. An agent of level $k = 1$ ("*L1* agent") 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* agent 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 $L(k-1)$, suggesting the interpretation that agents try to "outguess" their opponents.[3] To illustrate, consider a seminal game in this literature,[4] where players pick a number between 0 and 100 and the one whose number is closest to some

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

[2] Propositions 1, 2.1, 3, and 4 hold in the cognitive hierarchy model.

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

[4] This guessing game is used by Nagel (1995) to explain the *Lk* model. It was also mentioned in e.g. Moulin (1986) and Simonsen (1988).

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 k increases, the best response of *Lk* approaches 0, the only Nash equilibrium of the game.

This paper applies the *Lk* model to the AGV mechanism with one-dimensional types. We look at the case where the principal knows the type distribution and expects equilibrium behavior on part of the agents. Such principal is ignorant of the fact that he operates in an *Lk* environment. In this setting we conduct a positive exercise and find conditions under which the mechanism remains robust to *Lk*. Throughout the paper we assume 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 irrespective of the levels prevailing in the population.

Further, in the random-*L0* specification of *Lk*, we show that if the distribution of random actions (*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 interesting case where the type distribution used by the planner to assign transfers differs from *L1*s' expectation of the opponents' actions. 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 produces a sub-optimal allocation. Denoting the distribution of random *L0* strategies by Φ and the distribution of types by F, we study how the relation between Φ and F affects the *Lk* strategies in the mechanism.

We focus on the case where Φ dominates F (in the sense of first-order stochastic dominance) or vice versa. This corresponds to scenarios where players believe a salient strategy is to systematically under- or over-report one's type. The main result characterizes the deviations from equilibrium behavior for the case that the efficient choice rule is linear in agents' types (the environment we call neutral). If *L0* agents are expected to under-report their types, then all types of an *L1* agent will over-report their types to the mechanism, and vice versa. Therefore *L1* agents display compensatory bias in reports. The distortion carries over to higher levels, but the expected absolute value of the distortion of type decreases as level k goes up; in the case of quadratic utilities, the rate of decrease is exponential. Interestingly, the direction of the 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.

The results extend partially to the non-neutral case where types are complements or substitutes with respect to the efficient choice of allocation. Non-neutrality means that the marginal effect of one agent's type on the efficient allocation is *not* invariant in the

[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).

other agent's type. In particular, when the other agent's type is high, the marginal effect is stronger in case of complements and weaker in case of substitutes. In either of these environments reports have two counter-veiling effects on the choice of allocation. The first direct effect of compensating bias pushes the allocation in the direction of marginal payoff increase. The second indirect effect changes the choice rule's sensitivity to the opponent's report. Therefore, compensating bias remains best reply in type ranges where the direct effect dominates. We demonstrate by means of example that the dominance of the indirect effect changes the prediction.

While the main interest of this paper is positive, we conduct a separate normative analysis of the AGV mechanism. This part is concerned with a principal who is *aware* of the *Lk* environment and seeks the appropriate AGV-type mechanism for efficient implementation. In particular, we change the transfer rule to reflect the actual expected externality (under the level-k strategy profile) and thus to elicit the information correctly.[6] The *Lk* environment is characterized by three components: type distribution F, random actions distribution Φ and agents' levels k. When all three components are known, the efficient *Lk* mechanism differs from the original AGV in its transfer to *L1* agents only. By correcting the incentives at level 1 the principal restores truth-telling at all levels and achieves efficiency. When the information on F, Φ or k is missing, the principal can expand the mechanism to elicit the agents' knowledge. One way to do this is to add a betting round where the agents guess each others' reports. Ex post, the principal rewards correct guesses. Betting is a powerful tool for the elicitation of correlated information[7] and turns out to be instrumental in the *Lk* environment. We show how betting can be used to elicit levels k and other information necessary to construct the efficient mechanism.

This paper is among the first studies of mechanisms in an *Lk* environment. Crawford (2015) looks at the double auction mechanism and revisits Myerson and Satterthwaite's (1983) impossibility result in the *Lk* framework. He finds, in particular, that revelation principle does not hold in this framework since the choice of mechanism influences the correctness of *Lk* beliefs. Similar to his paper, the normative part of our analysis exploits the *predictably incorrect* beliefs of *Lk* agents. De Clippel et al. (2014) provide a characterization of implementable choice functions in a general setup with finite depth of reasoning. They consider the expected externality mechanism as an example and show that it achieves efficient implementation under the assumption that *L0* report truthfully. In contrast, the present paper allows for *L0* to be random and arbitrarily far from truthtelling.

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 shows how the AGV mechanism can be adjusted to the *Lk* environment, and Sect. 5 concludes.

[6] This was pointed out by an anonymous referee.

[7] See Myerson (1981), Crémer and McLean (1985, 1988).

2 The model

Preferences The preference environment is characterized by the following assumptions:

$A1$ Utilities are linear in money.
$A2$ Values are private.
$A3$ Values are independent draws from a commonly known distribution F with density f.

Assumptions $A1$ and $A2$ imply that the utility function of a given agent $i \in I = \{1, 2 \ldots n\}, n \geq 2$, can be represented as:

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

where $v_i(x, \theta_i)$ is the utility derived from allocation $x \in X \subseteq \mathcal{R}$, θ_i is the privately known preference parameter that we refer to as the agent's *type*, and T_i is the monetary transfer to agent i. Agent types θ_i are drawn independently from Θ, a compact subset of $\mathcal{R}$, according to a distribution F. We assume that $v_i(x, \theta_i)$ is strictly concave in x and continuously differentiable with respect to both arguments on the entire domain. Some of our results require that the preferences satisfy a single crossing (Spence–Mirrlees) condition. The condition postulates that the cross-derivative of $v_i(x, \theta_i)$ with respect to allocation x and type θ_i has constant sign over the function's domain:

$A4$. $v_i(x, \theta_i)$ satisfies the Spence–Mirrlees condition, i.e., either $A4.1$ or $A4.2$ holds:

$A4.1$ $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x, \theta_i) > 0$, for all i and $(x, \theta_i) \in (X, \Theta)$,

$A4.2$ $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x, \theta_i) < 0$, for all i and $(x, \theta_i) \in (X, \Theta)$.

$A1$-$A4$ are the basic assumptions of mechanism design. A further standard assumption is that agents play Bayes–Nash equilibrium: the profile of strategies is a fixed point of a best reply correspondence. In this paper, we consider a framework with a finite number of best-reply iterations that do not generally start at equilibrium. This framework 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 agent $i \in I$ of type θ_i and strategy profile $s = (s_1, s_2, \ldots s_n)$, where $s_i(\theta_i)$, or simply s_i, maps into an action. We look at agents *who engage in iterations of best reply*. The *Lk* strategy $s_i^{(k)}(\theta_i)$ is recursively defined as function of agent's type θ_i that maximizes his expected payoff against level-$(k-1)$ profile $s_{-i}^{(k-1)}(\theta_{-i})$. The agent believes with certainty that his opponents make exactly $k-1$ iterations of best reply.[8] As starting point of the recursion, the model features nonstrategic *L0* agents whose actions $s_i^{(0)}$ are drawn from a given distribution Φ. By analogy, we say that $s_i^{(0)}(\theta_i) \equiv s_i^{(0)}$ is an unobserved *random* mapping such that the induced cumulative distribution of actions is Φ and the density is φ.

[8] In contrast, *Cognitive hierarchy* model assumes that *Lk* agents attributes strictly positive probabilities to *all* the levels of rationality lower than k.

Definition For $k \geq 1$ the optimal strategy $s_i^{(k)}$ maximizes the expected payoff of agent i against $s_{-i}^{(k-1)}$:[9]

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

where θ_{-i} is the residual profile of types. The expectation is taken over the residual types and mappings $s_i^{(0)}$. The following simple observation establishes the relation between the Lk and equilibrium strategy profiles.[10]

Observation: If $s^{(k)}(\theta) = s^{(k+1)}(\theta)$ for some $k \geq 1$ and $\theta \in \Theta$, then $s^{(k)}(\theta)$ constitutes a Bayes-Nash equilibrium.

Choice rules and mechanisms For a quasi-linear utility representation (1), we define a choice rule $x^*(\theta)$ as *efficient* if it maximizes the total welfare for every profile of agents' types $\theta = (\theta_1, \theta_2, \ldots \theta_n)$:

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

We look at a direct mechanism, where the agents report their types to the principal: i's report s_i is a member of Θ.[11] A mechanism *implements* the choice rule $x^*(\cdot)$ if the profile of truth-telling reports is an equilibrium. The expected externality mechanism introduced in d' Aspremont and Gérard–Varet (AGV, 1979) is an example of such mechanism. AGV chooses the efficient allocation $x^*(\cdot)$ and assigns the following monetary transfers to the participants:

$$T_i(s) = t_i(s_i) - \frac{1}{n-1}\sum_{l \neq i} t_l(s_l), \tag{4}$$

where

$$t_i(s_i) = \mathbb{E}\sum_{j \neq i} v_j\left(x^*(s_i, \theta_{-i}); \theta_j\right). \tag{5}$$

The transfer $t_i(s_i)$ is constructed such that agent i internalizes the expected effect of his report on the others' welfare, assuming they tell the truth. This guarantees that agent i's incentives are aligned with the total welfare maximization, therefore truth-telling is Bayes–Nash equilibrium. Note that this implies immediately that in the truthful-$L0$ specification of the Lk model efficient implementation obtains for any k.

The second part of the transfer, $\frac{1}{n-1}\sum_{l \neq i} t_l(s_l)$, guarantees that mechanism satisfies ex post budget balance. In particular, in the level-k model the transfers add up to

[9] We consider problems where the solution is unique.

[10] The observation follows immediately from the definition of the Bayes–Nash equilibrium as fixed point of the best-reply correspondence (2).

[11] Generally, the revelation principle may fail in Lk environments, such that the restriction to direct mechanisms is not without loss (see Crawford 2015). In particular, the space of admissible messages may affect the beliefs of $L1$ players and consequently their best response in the mechanism.

zero after any profile of reports s.[12] Note that this part of transfer does not depend on i's own report s_i, therefore it can be omitted from the analysis of incentives.

Level-k in the Mechanism In the expected externality mechanism, an Lk agent, $k \geq 1$, maximizes the expected gain in the mechanism:

$$\mathbb{E}\left[v_i\left(x^*\left(s_i, s_{-i}^{(k-1)}\left(\theta_{-i}\right)\right); \theta_i\right) + t_i\left(s_i\right)\right] \tag{6}$$

Given the incentive transfer (5), the optimal Lk strategy in the mechanism is defined by the following:[13]

$$s_i^{(k)}\left(\theta_i\right) = \arg\max_{s_i \in \Theta} \mathbb{E}\left[v_i\left(x^*\left(s_i, s_{-i}^{(k-1)}\left(\theta_{-i}\right)\right); \theta_i\right) + \sum_{j \neq i} v_j\left(x^*\left(s_i, \theta_{-i}\right); \theta_j\right)\right] \tag{7}$$

Recall that a strategy profile that satisfies $s^{(k)}(\theta) = s^{(k-1)}(\theta)$ for all k and θ is a Bayes-Nash equilibrium. The following section demonstrates an example where this is not the case and studies the differences between Lk and equilibrium behavior in the AGV mechanism.

3 Unadjusted mechanism

This section takes the AGV mechanism as given and studies its outcomes in the Level-k environment. We establish the conditions under which the mechanism still yields efficient outcomes, and look at the misreporting of preferences that may arise in certain stochastic environments. We start with a simple example to illustrate some of our main findings.

Example Consider a setting with n agents 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 if the allocation departs from it. It is easy to verify that the socially efficient allocation is the average of individual bliss points: $x^*(\theta_1) = \frac{\sum_i \theta_i}{n}$. We prove the following simple lemma (see "Appendix").

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

$$s_i^{(k)}\left(\theta_i\right) = \theta_i + \Delta \times \left(-\frac{n-1}{n}\right)^k, \tag{8}$$

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 *agent.*

[12] In this respect, the AGV mechanism improves over the VCG mechanism (Vickrey, Clarke, and Groves), where ex post budget balance is generally impossible.

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

The *Lk* strategy (8) has several interesting properties. First, the size of the 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 pair of 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 $\Delta \times \left(\frac{n-1}{n}\right)^k$ between the true type θ and the *Lk* report $s_i^{(k)}(\theta_i)$ increases in the number of agents.

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

A5. $\frac{\partial^2 x^*}{\partial\theta_i \partial\theta_j}(\cdot) \equiv 0 \quad$ for all $i,\ j \in I$.

Level 1 is central to the entire analysis, since any distortion of truthtelling that emerges at *L1* propagates to higher levels. The analysis of *L1* optimal strategy:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in \Theta} \left\{ \mathbb{E}_{s_{-i}^{(0)}} v_i \left(x^* \left(s_i, s_{-i}^{(0)} \right); \theta_i \right) + \sum_{j \neq i} \mathbb{E}_{\theta_{-i}} v_j \left(x^* \left(s_i, \theta_{-i} \right); \theta_j \right) \right\} \tag{9}$$

yields the following proposition.

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

Proposition 1 establishes the equivalence between equilibrium and *Lk* predictions of the AGV mechanism's outcome. It shows that as long as the subjective distribution of random actions coincides with the (objective) distribution of types, it is irrelevant whether the agents stop at a finite level of reasoning or engage in equilibrium thinking. Proposition 1 trivially extends to the cognitive hierarchy (CH) model, since both *Lk* and CH models define level-*1* 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 $F \equiv \Phi$. Observe that the equivalence result does not rely on either the linearity of the social choice rule nor the Spence-Mirrlees condition.

If distributions F and Φ do not coincide, *Lk* agents do not report truthfully in general. To study the report biases, we concentrate on the case where F and Φ can be ordered with respect to first-order stochastic dominance relation, denoted $\succ_{FOSD}$. This corresponds to scenarios where players believe a salient strategy is to systematically under- or over-report one's type. We have the following result.

Proposition 2 *Under assumptions A1–A5, L1 agents distort their type reports upwards if* $F \succ_{FOSD} \Phi$, *and downwards if* $\Phi \succ_{FOSD} F$. *If either* $F \succ_{FOSD} \Phi$ *or* $\Phi \succ_{FOSD} F$, *then* $\lim_k \mathbb{E}_{\theta_i} \left| s_i^{(k)}(\theta_i) - \theta_i \right| = 0$, $\int\} \setminus \left(s_i^{(k)}(\theta_i) - \theta_i \right) = -\int\} \setminus \left(s_i^{(k-1)}(\theta_i) - \theta_i \right)$ *for all i.*

The proof of the proposition is given in the "Appendix". We start with the observation that any n-agent problem can be reduced to a problem with two agents due to

the fact that stochastic dominance is preserved under monotone transformations and summation of random variables. Then, in the framework with two agents, we analyze the first-order condition that corresponds to the payoff-maximization problem (9) to obtain the result.

The first part of Proposition 2 states that *L1* agents systematically (that is, for every realization of type) misreport their types, if one distribution dominates the other in the sense of first-order stochastic dominance. For example, if an *L1* agent expects *L0* agents' reports to dominate the type distribution, then *L1* will report a lower type than he actually has (and vice versa), even if this induces a less preferred allocation. The reason is that in the AGV mechanism, an agent's report affects both (1) the expected externality, which is calculated based on the true distribution F, and (2) the agent's own expected value from the allocation which depends on his own belief Φ about other agents' reports. If an agent believes the others over-report (Φ dominates), he concludes that the allocation is on average higher than it would be under truthful reports by the others. Given that the utility function is strictly concave, this reduces his perceived marginal value of the allocation, therefore he under-reports. If higher types prefer lower alternatives ('negative cross-derivative', as in *A*4.2), then *L0*s' over-reporting makes the chosen alternative lower and *L1* over-reports to compensate. In either case, an *L1* agent compensates the opponents' random behavior by misreporting his type in the opposite direction.

The second part of the proposition states that the expected deviation of reported from true types decreases in absolute value as the level of rationality increases. The sign of the expected deviation alternates at every transition from k to $k+1$. Thus the optimal level-k strategies follow a pattern similar to the example of Sect. 2. If level-*2* agents overstate their type in the game, then level-*3* agents 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* agents, then the expected chosen alternative is closer to efficiency.

Non-neutrality

The assumption of neutrality implies that the marginal effect of an agent's type on the efficient allocation is invariant in other agents' types. However, there are examples of preferences where this assumption is violated. Consider the case with two agents whose preferences are given by $v_1 = \theta_1 x$ for Agent 1 and $v_2 = -\frac{x^2}{2\theta_2}(\theta_2 > 0)$ for Agent 2. The optimal allocation is $x^* = \theta_1\theta_2$. Agent 1's utility in mechanism (excluding the budget balancing part)[14] equals $v_1 + t_1 = \mathbb{E}\left[\theta_1 x^*\left(\hat{\theta}_1, s_2^{(0)}\right) - \frac{\left(x^*\left(\hat{\theta}_1,\theta_2\right)\right)^2}{\theta_2}\right] = \theta_1\hat{\theta}_1\mathbb{E}s_2^{(0)} - (\hat{\theta}_1)^2\mathbb{E}\theta_2$. Suppose Φ dominates F such that $\mathbb{E}s_2^{(0)} = 1$ and $\mathbb{E}\theta_2 = 0$, then $v_1 + t_1 = \theta_1\hat{\theta}_1$. Thus Agent 1 will over-report if $\theta_1 > 0$ and under-report if $\theta_1 < 0$, which is not the prediction of Proposition 2. Contrary to the neutral environment, where $\Phi \succ F$ would imply under-reporting by all types of an *L1* agent (Proposition 2), this example features types that are complements with respect to the optimal allocation:

[14] Recall that the budget-balancing term does not depend on the agent's own report.

$\frac{\partial^2 x^*}{\partial\theta_i \partial\theta_j} = 1 > 0$. In such environments, the result of Proposition 2 holds only for a subset of types, as we demonstrate below.

Agents' types are **complements**[15] with respect to the efficient rule $\frac{\partial^2 x^*}{\partial\theta_i \partial\theta_j} > 0$ for all $i \neq j$. Agents' types are **substitutes**[16] with respect to the efficient rule $\frac{\partial^2 x^*}{\partial\theta_i \partial\theta_j} < 0$ for all $i \neq j$. When types are substitutes, a higher type by agent i lowers the marginal effect of the opponent's type. If types are complements, the interaction is the opposite: the marginal effect of j's type increases with the type of agent i.

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*). It says that the ratio of probability distribution functions $\frac{f(t)}{\varphi(t)}$ decreases in t if $\Phi \succ_{FOSD} F$, and increases in t if $F \succ_{FOSD} \Phi$.

Proposition 3 (a) *Under A1–A4.1, MLRP and complements environment,* $\exists t_i^*$ *such that for all types* $\theta_i < t_i^*$ *of* L1 *agent i he distorts his report downwards if* $\Phi \succ F$ *and upwards if* $F \succ \Phi$.

(b) *Under A1–A4.1, MLRP and substitutes environment,* $\exists t_i^*$ *such that for all types* $\theta_i > t_i^*$ *of* L1 *agent i he distorts his report downwards if* $\Phi \succ F$ *and upwards if* $F \succ \Phi$.

Proposition 4 (a) *Under A1–A4.2, MLRP and complements environment,* $\exists t_i^*$ *such that for all types* $\theta_i > t_i^*$ *of* L1 *agent i he distorts his report downwards if* $\Phi \succ F$ *and upwards if* $F \succ \Phi$.

(b) *Under A1–A4.2, MLRP and substitutes environment,* $\exists t_i^*$ *such that for all types* $\theta_i < t_i^*$ *of* L1 *agent i he distorts his report downwards if* $\Phi \succ F$ *and upwards if* $F \succ \Phi$.

Propositions 3 and 4 make four distinct claims. Consider the first claim, for example: 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 agents will tend to misreport their type so as to compensate the bias in the other agent's report. This claim is the same as Proposition 2, except that it does not include a range of valuations above a threshold. If there is first-order stochastic dominance in distributions, in the neutral case, an *L1* displays compensating behavior: *L1* systematically under- or over-reports, regardless of whether his true type is high or low. However, in a non-neutral case this is different. Observe that when types are complements or substitutes the mechanism may become

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

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

more sensitive to *L0's* misreporting in the extreme ranges of *L1's* type when *L1* misreports. Therefore *L1*'s strategy of compensating report bias has a further indirect effect on the allocation choice. For this reason, both Propositions 3 and 4 include only the type ranges that correspond to low enough sensitivity of the social choice rule to the other agent's report. Types in the low-sensitivity regions display the compensating behavior, similar to our benchmark result in Proposition 2.

Intuitively, the exclusion of some types in Propositions 3 and 4 can be understood as follows. Consider the more intuitive case of positive single crossing (A4.1). Suppose *L1* agent's type is high, so he prefers a high level of public good, and *complements* environment. Then compensatory under-reporting makes the choice rule less responsive to the opponent's over-reporting and thus may lead to the allocation being too low for his preferences. On the other hand compensatory over-reporting makes the choice rule more responsive to the opponent's under-reporting and thus, again, may lead to the choice of allocation that is too low. Suppose now that the agent's type is low, so he prefers a low level of public good, and *substitutes* environment, as in the example given at the beginning of this section. In the example the choice rule does not respond to the opponent's under-reporting and thus, if the agent over-reports his type, he increases the probability that the project is undertaken, and that is against his private interest. Therefore, the reaction of the choice rule to the opponent's report determines whether the compensating bias is a profitable strategy.

4 Adjusting the mechanism[17]

Our analysis so far assumed that the principal is unaware of the *Lk* environment. In other words, the principal implements the allocation and transfers as if the agents were infinitely rational. But what if the principal *knows* that the agents conduct only a finite number of best-reply iterations? How can he adjust the mechanism and achieve efficiency in this case? This section discusses this question. The answer depends critically on the principal's information about the setting. If the characteristics of stochastic setting—the type distribution F, distribution of random actions Φ, and the *Lk* identity of every agent—are known, then the principal can achieve efficiency by adjusting the incentive transfer. However, if some of that information is missing, the principal should expand the mechanism.

4.1 Known environment (F, Φ, k)

When F, Φ, and k_i for all $i \in I$ are known, the principal's response to the *Lk* environment is to adjust the incentive transfers accordingly. Knowing that *L1* agents expect their opponents to behave non-strategically according to the distribution Φ, the principal assigns the following transfer to any *L1* agent:

$$t_i^{(1)}(s_i) = \mathbb{E}_{s_{-i}^{(0)}} \sum_{j \neq i} v_j \left(x^* \left(s_i, s_{-i}^{(0)} \right); s_j^{(0)} \right) \tag{10}$$

[17] I am grateful to the anonymous referee who suggested writing this section and offered some important insights into adjusting the AGV mechanism.

The expectation in (10) is taken over the $L0$ strategies $s_{-i}^{(0)}$, as opposed to type distributions as in the original AGV mechanism.

Thus, the incentive transfer to all higher-level agents Lk remains unchanged relative to the original AGV mechanism:

$$t_i^{(>1)}(s_i) = \mathbb{E}_{\theta_{-i}} \sum_{j \neq i} v_j \left(x^* (s_i, \theta_{-i}) ; \theta_j \right) \tag{11}$$

Let AGV$k(F, \Phi)$ refer to the AGV mechanism with transfers Eqs. (10) and (11).

Lemma 2 *Any* Lk *player* ($k \geq 1$) *is truthful in AGV*k (F, Φ).

Proof Facing transfer (10), any *L1* agents report their types truthfully, since $s_i = \theta_i$ solves the utility maximization problem:

$$\max_{s_i \in \Theta} \mathbb{E}_{s_{-i}^{(0)}} \left[v_i \left(x^* \left(s_i, s_{-i}^{(0)} \right) ; \theta_i \right) + \sum_{j \neq i} v_j \left(x^* \left(s_i, s_{-i}^{(0)} \right) ; s_j^{(0)} \right) \right]. \tag{12}$$

Provided that *L1*s receive transfers that make them reveal their types, *L2*s hold a belief over the reports that coincides with F, the distribution of types. Similar to the Bayes Nash equilibrium in the standard AGV mechanism *L2* best replies to the incentives by reporting his type truthfully. By induction, truthfulness extends to all subsequent levels that face the standard AGV transfer (11). The induction relies on the fact that $L(k+1)$ believe that Lk best reply to $L(k-1)$ and believe that $L(k-1)$ best reply to $L(k-2)$ etc up to *L1*. □

Therefore, in case where the stochastic Lk environment is known, the principal can implement the efficient allocation by changing the transfer to *L1* agents only. As before, budget balance ex post is achieved through an additional term that is independent of agent i's own report s_i : $T_i(s) = t_i(s_i) - \frac{1}{n-1} \sum_{j \neq i} t_j \left(s_j \right)$.

4.2 Unknown environment

The construction of transfers Eqs. (10) and (11) relies on the principal's knowledge of distributions Φ and F, respectively. The *assignment* of transfers to agents relies on the knowledge of levels k_i for $i \in I$. If any part of this information is not available to the principal he has to elicit it from the agents. Unfortunately, there is little hope to get the information "for free". Suppose that the principal knew he was facing an *L1* agent i and asked him to report Φ. The agent would benefit from misrepresenting Φ as it determines his incentive transfer (10). For example, in the quadratic utility case (Sect. 3) the agent gains in Φ-expected externality if Φ is such that the other agents' preferences are very similar to his own preference report $\hat{\theta}_i$. In the extreme case, the agent reports a degenerate distribution Φ with a mass point at $\hat{\theta}_i$. Asking an *L2* agent to report Φ would not result in truthful elicitation either. Contrary to *L1*, misreporting Φ does not affect *L2*'s incentive transfer, but it does affect his expectation of the resulting

allocation choice. Since an *L2* believes that others are *L1* he also believes that their type reports can be manipulated by falsely reporting Φ. Furthermore, since *L2* believes that he pays a fraction $\frac{1}{n-1}$ of *L1*s' total incentive transfers as part of the budget balance program, his report of Φ also affects his monetary gain in the mechanism. These considerations illustrate the need for a proper elicitation mechanism.

Let P_i denote agent i's true belief about $(i+1)$'s moves[18] and $\hat{P}_i$ denote the reported belief. We assume that beliefs are differentiable for simplicity. Observe that $P_i = \Phi$, if $k_i = 1$. However if $k_i \geq 2$ then $P_i = F$ under the assumption of truth-telling Lk. Neither F, $\mathbf{\Phi}$ or levels k are known to the principal.

Consider the following two-stage AGVk (**TS-AGVk**) mechanism:

Stage 1 Agent i reports $\hat{P}_i$.[19]
Stage 2 First-stage reports pin down the transfer schedule and i reports type $\hat{\theta}_i$.

The principal implements the efficient allocation (3) and pays the transfer:[20]

$$t_i + b_i - \frac{1}{n-1} \sum_{l \neq i} t_l - b_{i+1}, \tag{13}$$

where $t_i = t_i(s_i) = \mathbb{E} \sum_{j \neq i} v_j \left(x^*(s_i, s_{-i}); s_j\right)$, expectation over s_{-i} is taken w.r.t. $\hat{P}_i$ (incentive part); $b_i = b_i\left(\hat{p}_i(s_{i+1})\right) = \lambda \ln \hat{p}_i(s_{i+1})$ (proper scoring or betting part), λ is a scalar and $\hat{p}_i(s_{i+1}) = \frac{\partial}{\partial s_{i+1}} \hat{P}_i(s_{i+1})$.[21] Note that compared to the standard AGV mechanism, the budget balancing part in TS-AGVk includes an extra term $-b_{i+1}$ to balance the betting rewards.

Lemma 3 *For any $\varepsilon > 0$ there exists $\lambda > 0$ in the TS-AGV*k *mechanism with $n > 2$, such that truth-telling is ε-optimal for an Lk-agent, given that I/i tell the truth.*[22]

Under the assumption that all agents tell the truth, the lemma states that no Lk-agent can deviate and gain more than ε by lying to the principal if the betting transfer is appropriately scaled. The proof is given in the "Appendix". The proof relies on the observation that the expected betting transfer: $\mathbb{E}b_i\left(s_{i+1}; \hat{p}_i\right) = \lambda \int_\Theta \ln \hat{p}_i(s_{i+1})\, dP_i(s_{i+1})$ is maximized at $\hat{p}_i \equiv p_i$ (Good 1952). However, since the report $\hat{p}_i$ also affects i's incentive transfer $t_i(\cdot)$, the loss in betting reward has to be sufficiently large to nullify any gain from changing the allocation and $t_i(\cdot)$ that i may achieve by misreporting p_i and θ_i.

Remark TS-AGVk does not rely on the knowledge that the underlying model is Lk. Specifically, the transfers are constructed to induce truth-telling as best response of an

[18] If $i = n$, consider his beliefs about agent 1.

[19] Since communicating the entire distribution function may not seem tractable, assume that the distributions belong to a known parametric class. In that case, the agents have to communicate only a finite number of parameters. See, e.g., Brooks (2013) and Azar et al. (2012).

[20] If $i = n$ read "$t_n + b_n - \frac{1}{n-1} \sum_{l \neq n} t_l - b_1$".

[21] If $i = n$ read "$\hat{p}_n(s_1) = \frac{\partial}{\partial s_1} \hat{P}_n(s_1)$".

[22] In the standard setting, this corresponds to an *ε-equilibrium*.

agent with arbitrary beliefs, not necessary an Lk agent. In contrast, the mechanisms introduced below are tailored to the particular setting of Lk and are therefore less robust to the change of environment.[23]

If **F and Φ are known but levels *k* are unknown**, then the first stage of the mechanism above can be simplified. Here, we use the fact that in the Lk model, agent i's level k_i can be inferred from his belief about another agents' level k_j, $j \neq i$. At the first stage of **TS-AGV*k*(F, Φ)** the principal asks each agent to guess the level of another participant. To fix ideas, let agent 1 report on k_2, agent 2 reports on k_3, and so on until agent n who reports on k_1. In the Lk model, agent i's report $\hat{k}^i_{i+1}$ about agent $(i+1)$'s level is truthful, if it is just below the agent's own level: $\hat{k}^i_{i+1} = k_i - 1$. The true belief may not be correct (i.e., $\hat{k}^i_{i+1}$ may or may not equal k^i_{i+1}); moreover, at least one agent's belief must be incorrect.

The structure of transfers in TS-AGV$k(F, \Phi)$ is given by (13), where the incentive part t_i is given by (10), if $\hat{k}^i_{i+1} = 0$, and (11), if $\hat{k}^i_{i+1} \geq 1$; the betting transfer $b_i = b_i\left(\hat{k}^i_{i+1}\right)$ is 0, if $\hat{k}^i_{i+1} = k^i_{i+1}$, and $-\lambda$ otherwise.

Lemma 4 *There exists $\lambda > 0$ in TS-AGV*k(F, Φ) *with $n > 2$, such that truth-telling is Lk-optimal for agent $i \in I$, given that I/i tell the truth.*

Unlike the TS-AGVk mechanism, TS-AGV$k(F, \Phi)$ with the appropriately chosen "punishment level" λ induces *exact* truth-telling. This is achieved because the reported levels k take on only discrete values $(0, 1, 2, \ldots)$.

If **F and k are known but Φ is unknown**, then we can exploit the fact that Φ is common knowledge among the agents. The principal can use a shoot-the-liar protocol by asking the agents to report Φ and punishing them if there is no unanimity. In this mechanism, reporting Φ truthfully is best reply to the residual profile of truthful reports. However, truth-telling is not a unique solution. Establishing uniqueness could involve using "nuisance" strategies, as in Maskin (1985), or additional stages, as in Moore and Repullo (1988).

5 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.[24] 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 of type space 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

[23] I thank an anonymous referee for this remark.

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

paper, first, studies one of the most influential existing mechanisms, d'Aspremont and Gerard-Varet (1979), in the *Lk* environment.

The AGV mechanism implements the efficient choice rule in Bayes-Nash equilibrium. 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.[25] On the downside, the 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 *Lk* model, this is not entirely innocuous.

Using the setup of the *Lk* model we start by conducting a positive analysis of the mechanism in the behavioral environment. We show that if there is a systematic difference in the perceptions of random-L0 actions and true types, then the agents distort their types at the first level and, by extension, also at the higher levels of rationality. Thereby we observe compensating behavior of finite-level agents in an AGV mechanism, that is, distorting one's report in the opposite direction to the opponents' anticipated bias. 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 actions, as opposed to types, to achieve truth-telling among *Lk* agents. Consequently, we adjust the AGV mechanism by changing transfer for *L1* agents in the case where the principal's has sufficient information. Otherwise, we introduce a betting scheme to elicit the agents' knowledge of the environment that the principal uses at a subsequent stage to induce truth-telling.

Altogether, our results suggest that the AGV mechanism is fairly robust to the iterative thinking environment. First, in the truthful-*L0* specification there is no distortion of truth-telling 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, the mechanism can be adjusted to the *Lk* framework in a way that maintains its key properties.

[25] Dominant-strategy implementation guarantees that the VCG mechanism achieves truthful revelation and efficiency in the *Lk* model ($k > 0$).

Appendix

Lemma 1

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

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

$$s^{(k-1)}\left(\theta_j\right) = \theta_i + \left(-\frac{n-1}{n}\right)^{k-1} \Delta \tag{14}$$

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

$$\begin{aligned} s_i^{(k)}(\theta_i) = \arg\max_{s_i \in \Theta} \mathbb{E}_{\theta_{-i}} & \left[\theta_i \left(\frac{s_i + \sum_{j \neq i} s^{(k-1)}\left(\theta_j\right)}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} s^{(k-1)}\left(\theta_j\right)}{n} \right)^2 \right. \\ & \left. + \sum_{j \neq i} \mathbb{E}_{\theta_{-i}} \left[\theta_{-i} \left(\frac{s_i + \sum_{j \neq i} \theta_{-i}}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} \theta_{-i}}{n} \right)^2 \right] \right] \\ = \theta_i + \frac{n-1}{n} & \left(\mathbb{E}\theta_j - \mathbb{E} s^{(k-1)}\left(\theta_j\right) \right) \qquad (15) \\ = \theta_i + \frac{n-1}{n} & \left(\mathbb{E}\theta_j - \mathbb{E}\left[\theta_j + \left(-\frac{n-1}{n} \right)^{k-1} \Delta \right] \right) = \theta_i + \left(-\frac{n-1}{n} \right)^k \Delta \qquad (16) \end{aligned}$$

Thus, if (14) holds on level $k-1$ it also holds on level k. Level-1 strategy is best reply to the profile of random actions:

$$\begin{aligned} s_i^{(1)}(\theta_i) = \arg\max_{s_i \in \Theta} \mathbb{E}_{s_{-i}^{(0)}} & \left[\theta_i \left(\frac{s_i + \sum_{j \neq i} s_j^{(0)}}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} s_j^{(0)}}{n} \right)^2 \right. \\ & \left. + \sum_{j \neq i} \mathbb{E}_{\theta_{-i}} \left[\theta_j \left(\frac{s_i + \sum_{j \neq i} \theta_j}{n} \right) - \frac{1}{2} \left(\frac{s_i + \sum_{j \neq i} \theta_j}{n} \right)^2 \right] \right] \\ = \theta_i - \frac{n-1}{n} \Delta, & \qquad (17) \end{aligned}$$

Thus for *L1* the induction formula (14) applies. □

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 (9) is the following:

$$\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(0)}\right); \theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(0)}\right)\right] + \sum_{j\neq i}\mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_j}{\partial x}(x^*(s_i, \theta_{-i}); \theta_j)\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})\right] = 0 \tag{18}$$

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

$$\sum_{j\neq i}\frac{\partial v_j}{\partial x}(x^*(s_i, \theta_{-i}); \theta_j) + \frac{\partial v_i}{\partial x}(x^*(s_i, \theta_{-i}); s_i) = 0. \tag{19}$$

Then the second term of (18) can be rewritten, such that the f.o.c. becomes:[26]

$$\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(0)}\right); \theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(0)}\right)\right] - \mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}\left(x\left(s_i, \theta_{-i}\right); s_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, \theta_{-i}\right)\right] = 0 \tag{20}$$

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 (20) and thus $s_i^{(1)}(\theta_i) = \theta_i$. □

Lemma A *Let us denote the following* L1 *maximization problem with n agents by* P_n*:*

$$\max_{s_i\in\Theta}\mathbb{E}\left[v_i\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right); \theta_i\right) + \sum_{j\neq i}v_j\left(x^*\left(s_i, \theta_{-i}\right); \theta_j\right)\right] \tag{21}$$

Statement Suppose that A1–A5 hold. Consider an *L1* problem P_n with n agents and $F \prec_{FOSD} \Phi$ ($\Phi \prec_{FOSD} F$). There exists an *L1* problem P_2 with 2 agents 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 $\frac{\partial^2 x^*}{\partial s_i \partial s_j} \equiv 0$ (A5) implies that $x^*(s_1, \ldots s_n) = \sum_i \lambda_i h_i(s_i)$ for some scalars $\lambda_i, \lambda_i > 0$ and monotone functions h_i. Without loss of generality,

[26] 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).

consider $h_i(s_i) \equiv s_i$. Condition (20) can be rewritten as follows:

$$\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(\sum_{j\neq i}\lambda_j s_j^{(0)}+\lambda_i s_i;\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)\right]$$
$$=\mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}\left(\sum_{j\neq i}\lambda_j\theta_j+\lambda_i s_i;s_i\right)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})\right]. \tag{22}$$

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 agent in P_2. In this problem P_2 the first order condition writes as follows:

$$\mathbb{E}_{s_\Sigma^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(s_\Sigma^{(0)}+\lambda_i s_i;\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_\Sigma^{(0)}\right)\right]=\mathbb{E}_{\theta_\Sigma}\left[\frac{\partial v_i}{\partial x}(\theta_\Sigma+\lambda_i s_i;s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_\Sigma)\right]. \tag{23}$$

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

Lemma B

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

$$s_i^{(1)}(\theta_i)=\theta_i+\frac{\int(F(t)-\Phi(t))\,d\frac{\partial v_i}{\partial x}\left(x^*\left(s_i^{(1)}(\theta_i),t\right);s_i^{(1)}(\theta_i)\right)\frac{\partial x^*}{\partial s_i}\left(s_i^{(1)}(\theta_i),t\right)}{\int\frac{\partial^2 v_i}{\partial x\,\partial\theta_i}\left(x^*\left(s_i^{(1)}(\theta_i),s_{-i}^{(0)}\right);\widehat{\theta}_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i^{(1)}(\theta_i),s_{-i}^{(0)}\right)d\Phi(s_{-i}^{(0)})} \tag{24}$$

Proof Rewrite (20) as follows:

$$0=\mathbb{E}_{s_{-i}^{(0)}}\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)\right]$$
$$-\mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})\right] \tag{25}$$
$$=\int\frac{\partial v_i}{\partial x}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)d\Phi(s_{-i}^{(0)})$$
$$-\int\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})dF(\theta_{-i}) \tag{26}$$

Integrate the second term of Equation (26) by parts:

$$\begin{aligned}
&\int \frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})dF(\theta_{-i}) \\
&= \frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})F(\theta_{-i})\Big|_{\Theta} \\
&\quad - \int F(\theta_{-i})d\frac{\partial v_i}{\partial x}(x^*(s_i,\theta_{-i});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})
\end{aligned} \tag{27}$$

Modify the first term of Equation (26) by taking Taylor expansion under the integral:

$$\begin{aligned}
&\int \frac{\partial v_i}{\partial x}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)d\Phi\left(s_{-i}^{(0)}\right) \\
&= \int\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i,s_{-i}^{(0)}\right);s_i\right)+\frac{\partial^2 v_i}{\partial x\partial\theta_i}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\widehat{\theta_i}\right)(\theta_i-s_i)\right] \\
&\quad\times\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)d\Phi\left(s_{-i}^{(0)}\right)
\end{aligned} \tag{28}$$

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

$$\begin{aligned}
&= \int \frac{\partial v_i}{\partial x}\left(x^*(s_i,s_{-i}^{(0)});s_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)d\Phi\left(s_{-i}^{(0)}\right) \\
&\quad+\int \frac{\partial^2 v_i}{\partial x\partial\theta_i}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\widehat{\theta_i}\right)(\theta_i-s_i)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)d\Phi\left(s_{-i}^{(0)}\right)=
\end{aligned} \tag{29}$$

and integrate by parts:

$$\begin{aligned}
&= \frac{\partial v_i}{\partial x}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)\Phi\left(s_{-i}^{(0)}\right)\Big|_{\Theta} \\
&\quad-\int\Phi(t)d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) \\
&\quad+\int \frac{\partial^2 v_i}{\partial x\partial\theta_i}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\widehat{\theta_i}\right)(\theta_i-s_i)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)d\Phi\left(s_{-i}^{(0)}\right)
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
&\frac{\partial v_i}{\partial x}\left(x^*\left(s_i,s_{-i}^{(0)}\right);\theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i,s_{-i}^{(0)}\right)\Phi\left(s_{-i}^{(0)}\right)\Big|_{\Theta} \\
&= \frac{\partial v_i}{\partial x}\left(x^*\left(s_i,\theta_{-i}\right);s_i\right)\frac{\partial x^*}{\partial s_i}(s_i,\theta_{-i})F(\theta_{-i})\Big|_{\Theta}
\end{aligned} \tag{31}$$

Thus, the f.o.c. becomes:

$$\int (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i, t); s_i)\frac{\partial x^*}{\partial s_i}(s_i, t) + (\theta_i - s_i)\int \frac{\partial^2 v_i}{\partial x \partial \theta_i}\left(x^*\left(s_i, s_{-i}^{(0)}\right); \widehat{\theta_i}\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(0)}\right) d\Phi\left(s_{-i}^{(0)}\right) = 0 \quad (32)$$

We can rewrite the solution as follows:

$$s_i^{(1)}(\theta_i) - \theta_i \equiv \frac{\int (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}\left(x^*\left(s_i^{(1)}(\theta_i), t\right); s_i^{(1)}(\theta_i)\right)\frac{\partial x^*}{\partial s_i}\left(s_i^{(1)}(\theta_i), t\right)}{\int \frac{\partial^2 v_i}{\partial x \partial \theta_i}\left(x^*\left(s_i^{(1)}(\theta_i), s_{-i}^{(0)}\right); \widehat{\theta_i}\right)\frac{\partial x^*}{\partial s_i}\left(s_i^{(1)}(\theta_i), s_{-i}^{(0)}\right) d\Phi(s_{-i}^{(0)})} \quad (33)$$

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

Lemma C

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

$$\frac{\partial^2 v_i}{\partial x \partial \theta_i}\left(x^*\left(s_i^{(1)}(\theta_i), s_{-i}^{(0)}\right); \widehat{\theta_i}\right)\frac{\partial x^*}{\partial s_i}\left(s_i^{(1)}(\theta_i), s_{-i}^{(0)}\right) > 0. \quad (34)$$

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

$$\frac{\partial v_i}{\partial x}(x^*(t_i, t_{-i}), t_i) + \frac{\partial v_{-i}}{\partial x}(x^*(t_i, t_{-i}), t_{-i}) \equiv 0 \quad (35)$$

Differentiate with respect to θ_i:

$$\frac{\partial x^*}{\partial s_i}(t_i, t_{-i})\left[\frac{\partial^2 v_i}{\partial x^2}(x^*(t_i, t_{-i}), t_i) + \frac{\partial^2 v_{-i}}{\partial x^2}(x^*(t_i, t_{-i}), t_{-i}\right] + \frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(t_i, t_{-i}), t_i) = 0 \quad (36)$$

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

$$\frac{\partial^2 v_i}{\partial x^2}\left(x^*(t_i, t_{-i}), t_i\right) + \frac{\partial^2 v_{-i}}{\partial x^2}\left(x^*(t_i, t_{-i}), t_{-i}\right) < 0 \quad (37)$$

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

$$\operatorname{sgn}\left(\frac{\partial x^*}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)})\right) = \operatorname{sgn}\left(\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i^{(1)}(\theta_i), s_{-i}^{(0)}), s_i^{(1)}(\theta_i))\right). \tag{38}$$

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. □

Proposition 2

Statement 2.1 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 Eq. (33). Lemma C (p. 25) shows that the denominator of the expression is positive. Let us transform the nominator as follows:

$$\begin{aligned}
&\int (F(t) - \Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i, t); s_i)\frac{\partial x^*}{\partial s_i}(s_i, t) \\
&= \int (F(t) - \Phi(t)) \Bigg[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, t); s_i)}_{-(1)} \underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i, t)\frac{\partial x^*}{\partial s_i}(s_i, t)}_{+(2)} \\
&\quad + \frac{\partial v_i}{\partial x}(x^*(s_i, t); s_i) \underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i, t)}_{=0(3)} \Bigg] dt
\end{aligned} \tag{39}$$

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. 25), $\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*; \theta_i)\frac{\partial x^*}{\partial s_i} > 0$ 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

$$\left[\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, t); s_i)\frac{\partial x^*}{\partial s_{-i}}(s_i, t)\frac{\partial x^*}{\partial s_i}(s_i, t) + \frac{\partial v_i}{\partial x}(x^*(s_i, t); s_i)\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i, t)\right] \tag{40}$$

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. □

Statement 2.2 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 $sgn\left(s_i^{(k)}(\theta_i) - \theta_i\right) = -sgn\left(s_i^{(k-1)}(\theta_i) - \theta_i\right)$.

Proof Recall that by definition:

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in \Theta} \mathbb{E}_{\theta_{-i}}\left[v_i\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right); \theta_i\right) + v_{-i}(x^*(s_i, \theta_{-i}); \theta_{-i})\right] \tag{41}$$

The first-order condition for level-k strategy $s_i^{(k)}(\theta_i)$ is as follows ($s_i^{(k)}(\theta_i) = s_i$):[27]

$$\begin{aligned}
0 &= \mathbb{E}_{\theta_{-i}}\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right); \theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right)\right. \\
&\quad \left. + \frac{\partial v_{-i}}{\partial x}(x^*(s_i, \theta_{-i}); \theta_{-i})\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})\right] \qquad (42)\\
&= \mathbb{E}_{\theta_{-i}}\left[\left[\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right); \theta_i\right)\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right)\right]\right. \\
&\quad \left. - \frac{\partial v_i}{\partial x}\left(x^*\left(s_i, \theta_{-i}\right); s_i\right)\frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})\right] \qquad (43)\\
&\overset{(*)}{=} \mathbb{E}_{\theta_{-i}}\left[\left(\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right); \theta_i\right) - \frac{\partial v_i}{\partial x}\left(x^*\left(s_i, \theta_{-i}\right); s_i\right)\right)\right. \\
&\quad \times \frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right) \\
&\quad \left. + \frac{\partial v_i}{\partial x}\left(x^*\left(s_i, \theta_{-i}\right); s_i\right)\underbrace{\left(\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right) - \frac{\partial x^*}{\partial s_i}(s_i, \theta_{-i})\right)}_{=0}\right]. \qquad (44)
\end{aligned}$$

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

$$\begin{aligned}
0 &= \mathbb{E}_{\theta_{-i}}\left[\left(\frac{\partial v_i}{\partial x}\left(x^*\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right); \theta_i\right) - \frac{\partial v_i}{\partial x}\left(x^*\left(s_i, \theta_{-i}\right); s_i\right)\right)\right. \\
&\quad \left. \times \frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right)\right] \qquad (45)\\
&= \mathbb{E}_{\theta_{-i}}\left[\frac{\partial^2 v_i}{\partial x^2}\left(x^*\left(s_i, \widehat{s_{-i}}\right); \widehat{\theta_i}\right)\frac{\partial x^*}{\partial s_i}\left(s_i, \widehat{s_{-i}}\right)\left(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i}\right)\right. \\
&\quad \left. + \frac{\partial^2 v_i}{\partial x \partial \theta_i}\left(x^*\left(s_i, \widehat{s_{-i}}\right); \widehat{\theta_i}\right)(\theta_i - s_i)\right]\frac{\partial x^*}{\partial s_i}\left(s_i, s_{-i}^{(k-1)}(\theta_{-i})\right) \qquad (46)
\end{aligned}$$

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

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

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

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

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

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

$$\frac{-\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_i) \frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}}) - \frac{\partial^2 v_{-i}}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_{-i}) \frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); s_i)} = 1, \quad (47)$$

thus $\frac{-\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); s_i) \frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); s_i)} < 1$.[28]

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

$$\frac{-\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \hat{\theta}_i) \frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \hat{\theta}_i)} < 1 \quad (48)$$

as well. Take the expectation of both sides:

$$\mathbb{E}_{\theta_i}\left[s_i^{(k)}(\theta_i) - \theta_i\right] = \mathbb{E}_{\theta_i}\mathbb{E}_{\theta_{-i}}\left[\frac{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i}) \frac{\partial x^*}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x \partial \theta_i}(x^*(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})\right] \quad (49)$$

as types are independent and the distributions of types coincide,

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

$$\mathbb{E}_{\theta_i}\left[s_i^{(k)}(\theta_i)-\theta_i\right]=\mathbb{E}_{\theta_{-i}}\left[(s_{-i}^{(k-1)}(\theta_{-i})-\theta_{-i})\mathbb{E}_{\theta_i}\frac{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,\widehat{s_{-i}});\widehat{\theta_i})\frac{\partial x^*}{\partial s_i}(s_i,\widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial x\partial\theta_i}(x^*(s_i,\widehat{s_{-i}});\widehat{\theta_i})}\right] \tag{50}$$

$$\mathbb{E}_{\theta_i}\left|s_i^{(k)}(\theta_i)-\theta_i\right|<\mathbb{E}_{\theta_i}\left|s_i^{(k-1)}(\theta_i)-\theta_i\right| \tag{51}$$

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 (51) 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 28). 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 51 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 $\overline{L}=0$).

This concludes the proof of Proposition 2. □

Proposition 3

Proposition 3a *Under A1–A4.1, MLRP and complements environment, $\exists t_i^*$ such that for all $\theta_i<t_i^*$ 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 3b *Under A1–A4.1, MLRP and substitutes environment, $\exists t_i^*$ such that for all $\theta_i>t_i^*$ 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 Eq. (33). Start with the case of Proposition 3a:

$\frac{\partial^2 v_i}{\partial x\partial\theta_i}(x,\theta_i)>0$, $\frac{\partial^2 x^*}{\partial s_i\partial s_{-i}}(s_i,t)\geq 0$. The nominator:

$$\begin{aligned}&\int_{\underline{t}}^{+\infty}(F(t)-\Phi(t))\,d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)\\&=\underbrace{\int_{s_i}^{+\infty}(F(t)-\Phi(t))\left[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-(1)}\underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)}_{+(2)}\underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+(2)}+\underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{-(3)}\underbrace{\frac{\partial^2 x^*}{\partial s_i\partial s_{-i}}(s_i,t)}_{+(4)}\right]dt}_{\text{“first term”}}\\&+\underbrace{\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)}_{\text{“second term”}}\end{aligned} \tag{52}$$

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[29] and $\frac{\partial v_i}{\partial x}(x^*(s_i,s_i);s_i)=0$. Consider *the first term* in brackets:

29 $\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial t}(s_i,t)<0$.

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.

Thus we obtain that

$$\left[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-}\underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)}_{+}\underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+} + \underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{-}\underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)}_{+}\right] \tag{53}$$

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

$$\begin{aligned}
&\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)\\
&=\underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{=0}\frac{\partial x^*}{\partial s_i}(s_i,s_i)\,(F(s_i)-\Phi(s_i))\\
&\quad+\frac{\partial v_i}{\partial x}(x^*(s_i,\underline{t});s_i)\frac{\partial x^*}{\partial s_i}(s_i,\underline{t})\underbrace{\big(F(\underline{t})-\Phi(\underline{t})\big)}_{=0}\\
&\quad-\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)d\,(F(t)-\Phi(t))\\
&=-\int_{\underline{t}}^{s_i}(F(t)-\Phi(t))\,\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t)\,(f(t)-\varphi(t))\,dt,
\end{aligned} \tag{54}$$

where

$$\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i) \geq 0, \tag{55}$$

for $t \leq s_i$

$$\frac{\partial x^*}{\partial s_i}(s_i,t) > 0. \tag{56}$$

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, due to the following. By the MLRP assumption, $\frac{f(t)}{\varphi(t)}$ decreases in t; thus, there exists a t_i^* such that $f(t_i^*)-\varphi(t_i^*)=0$. This implies that, for θ_i such that $s_i^{(1)}(\theta_i) \leq t_i^*$, the result is established: the *L1*s with sufficiently low types distort their 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 3a is now proven. □

To prove Proposition 3b ($\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t) \leq 0$), we change the decomposition of the nominator as follows:

$$\begin{aligned}
&\int_{\underline{t}}^{+\infty} (F(t)-\Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) \\
&= \int_{\underline{t}}^{s_i} (F(t)-\Phi(t)) \Bigg[\underbrace{\frac{\partial^2 v_i}{\partial x^2}(x^*(s_i,t);s_i)}_{-} \underbrace{\frac{\partial x^*}{\partial s_{-i}}(s_i,t)}_{+} \underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+} \\
&\quad + \underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{+} \underbrace{\frac{\partial^2 x^*}{\partial s_i \partial s_{-i}}(s_i,t)}_{-} \Bigg] dt \\
&\quad + \int_{s_i}^{+\infty} (F(t)-\Phi(t))\, d\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)\frac{\partial x^*}{\partial s_i}(s_i,t) \qquad (57)
\end{aligned}$$

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:

$$-\int_{s_i}^{+\infty} \underbrace{\frac{\partial v_i}{\partial x}(x^*(s_i,t);s_i)}_{-} \underbrace{\frac{\partial x^*}{\partial s_i}(s_i,t)}_{+} (f(t)-\varphi(t))\, dt. \qquad (58)$$

Similarly to the argument in 3a, we identify the condition under which both parts of the nominator have the same sign. Given the decomposition (57), we can see that for this to hold s_i has to be sufficiently *high* (or θ_i such that $s_i^{(1)}(\theta_i) \geq t_i^*$). Proposition 3b proven. □

Proof of Proposition 4 The statement and proof are symmetric to Proposition 3. □

Proof of Lemma 3 Fix an arbitrary $\varepsilon > 0$. The subjective expected gain in deviation from truthfully reporting (P_i,θ_i) to $(\hat{P}_i,\hat{\theta}_i)$ amounts to:

$$\begin{aligned}
D\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right) &\equiv \Delta W\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right) \\
&\quad + \lambda \int \left[\ln \hat{p}_i\,(s_{i+1}) - \ln p_i\,(s_{i+1})\right] p_i\,(s_{i+1})\, ds_{i+1}, \qquad (59)
\end{aligned}$$

where

$$\begin{aligned}\Delta W\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\equiv&\iint_{\Theta^{n-1}}\left[v_i\left(x^*\left(\hat{\theta}_i,s_{-i}\right);\theta_i\right)-v_i\left(x^*\left(\theta_i,s_{-i}\right);\theta_i\right)\right]\\&\times p_i\left(s_{-i}\right)d^{n-1}s_{-i}\\&+\iint_{\Theta^{n-1}}\sum_{j\neq i}v_j\left(x^*\left(\hat{\theta}_i,s_{-i}\right);s_j\right)\hat{p}_i\left(s_{-i}\right)d^{n-1}s_{-i}\\&-\iint_{\Theta^{n-1}}\sum_{j\neq i}v_j\left(x^*\left(\theta_i,s_{-i}\right);s_j\right)p_i\left(s_{-i}\right)d^{n-1}s_{-i}.\end{aligned}\tag{60}$$

The classic result of Good (1952) implies that

$$\int\left[\ln\hat{p}_i\left(s_{i+1}\right)-\ln p_i\left(s_{i+1}\right)\right]p_i\left(s_{i+1}\right)ds_{i+1}\leq 0.\tag{61}$$

Therefore, $D\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\geq\varepsilon$ only if $\Delta W\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\geq\varepsilon$. Consider set Π containing all $\hat{P}_i$, P_i such that for $\Delta W\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\geq\varepsilon$ for a least some $\left(\hat{\theta}_i,\theta_i\right)\in\Theta^2$ and assume that Π is non-empty. Then we can define

$$C=\max_{\left(\hat{P}_i,P_i\right)\in\Pi}\max_{\left(\hat{\theta}_i,\theta_i\right)\in\Theta^2}\Delta W\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\tag{62}$$

and

$$-c=\max_{\left(\hat{P}_i,P_i\right)\in\Pi}\int\left[\ln\hat{p}_i\left(s_{i+1}\right)-\ln p_i\left(s_{i+1}\right)\right]p_i\left(s_{i+1}\right)ds_{i+1}.\tag{63}$$

C is the greatest reward for misreporting within Π and $c>0$ is the lowest punishment (before scaling) for misreporting within Π. The total gain from deviation (59) is capped:

$$D\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\leq C-\lambda c,\tag{64}$$

hence one can always find $\lambda>0$ such that $C-\lambda c<0$. Thus $D\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)<0$ and the premise of non-empty Π is false for the given λ. We have shown that for all $\varepsilon>0$ there exists $\lambda>0$ such that there exists no $\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)$ such that $D\left(\hat{P}_i,\hat{\theta}_i;P_i,\theta_i\right)\geq\varepsilon$. □

Proof of Lemma 4 The subjective expected gain in deviation from truthfully reporting (k^i_{i+1}, θ_i) to $(\hat{k}^i_{i+1}, \hat{\theta}_i)$, where $\hat{k}^i_{i+1} \neq k^i_{i+1} = k_i - 1$, amounts to:

$$\begin{aligned} D\left(\hat{k}^i_{i+1}, \hat{\theta}_i; k^i_{i+1}, \theta_i\right) \equiv & \iint_{\Theta^{n-1}} \left[v_i\left(x^*\left(\hat{\theta}_i, s_{-i}\right); \theta_i\right) - v_i\left(x^*\left(\theta_i, s_{-i}\right); \theta_i\right)\right] \\ & \times p_i\left(s_{-i}\right) d^{n-1} s_{-i} \\ & + \iint_{\Theta^{n-1}} \sum_{j \neq i} v_j\left(x^*\left(\hat{\theta}_i, s_{-i}\right); s_j\right) \hat{p}_i\left(s_{-i}\right) d^{n-1} s_{-i} \\ & - \iint_{\Theta^{n-1}} \sum_{j \neq i} v_j\left(x^*\left(\theta_i, s_{-i}\right); s_j\right) p_i\left(s_{-i}\right) d^{n-1} s_{-i} \\ & - \lambda \end{aligned} \tag{65}$$

where $\hat{p}_i$ and p_i are implied by $\hat{k}^i_{i+1}$ and k^i_{i+1}, respectively ($\hat{k}_{i+1} = 0$ implies $\hat{p}_i \equiv \phi$ and $\hat{k}_{i+1} \geq 1$ implies $\hat{p}_i \equiv f$). For any $\hat{k}^i_{i+1} \neq k^i_{i+1}$, let $C\left(\hat{k}^i_{i+1}, k^i_{i+1}\right)$ be the maximal value of $D(\hat{k}^i_{i+1}, \hat{\theta}_i; k^i_{i+1}, \theta_i) + \tilde{\lambda}$, where the maximization is over $(\hat{\theta}_i, \theta_i) \in \Theta^2$. Then, for any $(\hat{\theta}_i, \theta_i) \in \Theta^2$

$$D\left(\hat{k}^i_{i+1}, \hat{\theta}_i; k^i_{i+1 i+1}, \theta_i\right) \leq C\left(\hat{k}^i_{i+1}, k^i_{i+1}\right) - \lambda. \tag{66}$$

Clearly, one can always find $\lambda_i(\hat{k}^i_{i+1}, k^i_{i+1}) > 0$ such that $D(\hat{k}^i_{i+1}, \hat{\theta}_i; k^i_{i+1}, \theta_i) < 0$ if $\lambda = \lambda_i(\hat{k}^i_{i+1}, k^i_{i+1})$. Let $\lambda = \max\limits_{i, \hat{k}^i_{i+1}, k^i_{i+1}} \{\lambda_i(\hat{k}^i_{i+1}, k^i_{i+1})\}$, then for all $\hat{k}^i_{i+1}, \hat{\theta}_i, k^i_{i+1}, \theta_i$ we obtain: $D(\hat{k}^i_{i+1}, \hat{\theta}_i; k^i_{i+1}, \theta_i) < 0$. □

References

Athey S, Segal I (2013) An efficient dynamic mechanism. Econometrica 81(6):2463–2485

Azar P, Chen J, Micali S (2012) Crowdsourced bayesian auctions. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM, New York, pp 236–248

Bergemann D, Morris S (2005) Robust mechanism design. Econometrica 73(6):1771–1813

Brooks B (2013) Surveying and selling: Belief and surplus extraction in auctions. University of Chicago (Unpublished paper)

Camerer CF, Ho T-H (2015) Chapter 10–Behavioral game theory experiments and modeling, vol. 4 of Handbook of Game Theory with Economic Applications. Elsevier, Amsterdam, pp 517–573

Camerer CF, Ho T-H, Chong J-K (2004) A cognitive hierarchy model of games. Q J Econ 119(3):861–898

Costa-Gomes M, Crawford VP, Broseta B (2001) Cognition and behavior in normal-form games: an experimental study. Econometrica 69(5):1193–1235

Costa-Gomes MA, Crawford VP (2006) Cognition and behavior in two-person guessing games: an experimental study. Am Econ Rev 96(5):1737–1768

Crawford V (2015) Efficient Mechanisms for Level-k Bilateral Trading. Working paper

Crawford V, Kugler T, Neeman Z, Pauzner A (2009) Behaviorally optimal auction design: examples and observations. J Eur Econ Assoc 7(2–3):377–387

Crawford VP, Iriberri N (2007) 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

D'Aspremont C, Gerard-Varet L-A (1979) Incentives and incomplete information. J Public Econ 11(1):25–45

De Clippel G, Saran R, Serrano R (2014) Mechanism design with bounded depth of reasoning and small modeling mistakes. Working paper
Good IJ (1952) Rational decisions. J R Stat Soc Ser B (Methodol):107–114
Kets W (2012) Bounded reasoning and higher-order uncertainty. Working paper
Maskin E (1985) The theory of implementation in Nash equilibrium: a survey. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) Social goals and social organization: volume in memory of Elisha Pazner. Cambridge University Press, pp 173–204
Mathevet L (2010) Supermodular mechanism design. Theor Econ 5(3):403–443
Moore J, Repullo R (1988) Subgame perfect implementation. Econometrica 56(5):1191–1220
Moulin H (1986) Game theory for the social sciences. Series: studies in game theory and mathematical economics, 2nd and Revised Edition. New York University Press, New York, NY, USA. ISBN 9780814754306
Myerson RB, Satterthwaite MA (1983) Efficient mechanisms for bilateral trading. J Econ Theory 29(2):265–281
Nagel R (1995) Unraveling in guessing games: an experimental study. Am Econ Rev 85(5):1313–1326
Shaked M, Shanthikumar G (2007) Stochastic orders. Springer, New York
Simonsen MH (1988) Rational expectations, game theory and inflationary inertia. Econ Evolv Complex Syst 5:205–241
Stahl DO, Wilson PW (1994) Experimental evidence on players' models of other players. J Econ Behav Organ 25(3):309–327