Korpela, Ville

Working Paper

All Deceptions Are Not Alike: Bayesian Mechanism Design with Social Norm Against Lying

Discussion paper, No. 95

Provided in Cooperation with:
Aboa Centre for Economics (ACE), Turku

Suggested Citation: Korpela, Ville (2014) : All Deceptions Are Not Alike: Bayesian Mechanism Design with Social Norm Against Lying, Discussion paper, No. 95, Aboa Centre for Economics (ACE), Turku

This Version is available at:
https://hdl.handle.net/10419/233311

Ville Korpela

All Deceptions Are Not Alike: Bayesian Mechanism Design with Social Norm Against Lying

Aboa Centre for Economics
Discussion paper No. 95
Turku 2014

The Aboa Centre for Economics is a joint initiative of the economics departments of the University of Turku and Åbo Akademi University.

ISSN 1796-3133

Printed in Uniprint
Turku 2014

Ville Korpela

All Deceptions Are Not Alike: Bayesian Mechanism Design with Social Norm Against Lying

Aboa Centre for Economics
Discussion paper No. 95
November 2014 (first draft September 2014)

ABSTRACT

We say that a society has a weak norm against lying if, all other things being equal, agents rather lie in such a way that they do not get caught. We show that if this is the case, and it usually is, then Bayesian monotonicity is no longer a constraint in implementation and all incentive compatible social choice functions are Bayesian implementable. In contrast to the previous literature our result derives from a refinement of the standard Bayes-Nash equilibrium that does not rely on any kind of intrinsic lying aversion on which the experimental evidence is mixed. In addition, it suggests that the so called "multiple equilibrium problem" may not be that severe.

JEL Classification: B41, C72, D78, D82

Keywords: Deception, Implementation, Incentive compatibility, Revelation principle, Social norms and conventions

Contact information

Ville Korpela
Department of Economics
University of Turku
FI-20014, Finland
Email: ville.korpela (at) utu.fi

Acknowledgements

I thank Juuso Välimäki, Hannu Vartiainen, Hannu Salonen, Hannu Nurmi, Nadine Chlass and participants of SING10 (Krakow), HECER Applied Microeconomics Seminar (Helsinki) and UECE Lisbon Meetings 2014: Game Theory and Applications (Lisbon) for comments and suggestions.

Why on earth would anybody want to write that down?

- Robert W. Rosenthal

1 Introduction

From the start mechanism design has relied heavily on one of its foundation stones - the revelation principle. This principle says, roughly speaking, that any social choice function which can be realized as a Bayes-Nash equilibrium of some mechanism, can also be realized as a truthful equilibrium of a direct mechanism. It is hard to say who should be credited for discovering this principle. According to Myerson (2008) it is one of those theorems that was found independently by several authors (Dasgupta et al., 1979; Harris and Townsend, 1981; Holmström, 1977; Myerson, 1973; Rosenthal, 1978), who where all building on the ideas of Gibbard (1973) and Aumann (1974). The reason why it is so hard to give priority in this matter was best stated by Rosenthal when he saw the result re-stated some years after his own discovery: "Why on earth would anybody want to write that down?" (Radner and Ray, 2003).

However, as old as the principle itself, is the observation that the associated direct mechanism may have other equilibria besides the truthful one, some of which were not present in the original mechanism (Palfrey, 1992; Palfrey and Srivastava, 1993; Feldman and Serrano, 2006). The problem, commonly known as *the multiple equilibrium problem*, is that for some reason or another, a reason outside the mechanism and beyond the control of the mechanism designer, one of these other equilibria may become focal and end up being played. Moreover, there does not seem to exist any consensus on how sever this problem really is. Does it make the revelation principle completely useless, does the principle simply need to be qualified somehow, or is it merely a minor hump in the road, not much of a practical significance. Consider the following extremely simple example.

Assume that a seller is auctioning one indivisible object to buyers with

independent and identically distributed valuations on the closed interval [0,1].[1] A bidding strategy $b_i : [0,1] \to [0,1]$ of buyer i gives his bid as a function of the true valuation. In a second-price sealed-bid auction any strategy profile $b = (b_1, \ldots, b_n)$, such that $b_i \equiv 0$ for all $i \in N \setminus \{j\}$ and $b_j \equiv 1$, is a Bayes-Nash equilibrium. Moreover, the revenue of the seller is 0 in this equilibrium. We are not saying that bidders can be expected to coordinate into this equilibrium - it is unlikely. Nevertheless, this simple example shows that there is a real problem. If we apply the revelation principle to find a revenue maximizing mechanism for the seller, it turns out that there are many possibilities, the first-price auction and the second-price sealed-bid auction for example (Myerson, 1981). However, the 0 revenue equilibrium is a problem only under the latter one. In fact, there is frequently only one equilibrium under the first-price auctions (Maskin and Riley, 2003; Lebrun, 2006), while in second-price sealed-bid auctions there are often equilibia in which some bidders act aggressively. This is what we meant by saying that the result might always need to be qualified somehow.

One possible way to approach this issue is to look what the extensive literature on dishonesty and deceitful behavior, which is obviously at the core of the revelation principle, can tell us. Unfortunately, all behavioral experiments that we are aware of suggest that a certain fraction of people behave in a predicted way while the rest do not, the fit being far from perfect (see, for example, Fischbacher and Föllmi-Heusi, 2013; Gneezy, 2005; Amir et al., 2008; Kartik and Hurkens, 2009a; Greene and Paxton, 2009). So what should this fraction be – 70%, 80%, 90%, or perhaps a utopian 99.99%. Quite frankly, it is against the basic nature of mechanism design to use this kind of results. We do not want the mechanism to realize the social choice function some of the times, rather we want the mechanism to always realize it, or at least to a very large degree, the failure being an extremely rare and exceptional case. In other words, we need to use something that is a truly invariant feature of human behavior, something that even the pure logic of

[1] This example would work also with a finite type space which is what we assume in this paper.

the situation suggest as necessary.

This is where social norms, the customs and conventions that govern behavior in societies, can be helpful. No one can deny that there is such thing as a social norm against lying.[2] This does not imply that people do not lie, and there is certainly ample evidence on the contrary (e.g. Hao and Houser, 2011), rather it only implies that, other things being equal, people prefer not to get caught. It would be very convenient if the mechanism designer could operate in a medieval society where people genuinely believed that God is watching them and that lying is bad *per se*.[3] However, in the secular world that we live in today, the mechanism designer can only rely on one thing: When it does not affect the material outcome people rather lie in such a way that they do not get caught. Surprisingly, as weak as this regularity may seems to be, it turns out that social norm against lying is sufficient to guarantee that there is nothing wrong in relying on the revelation principle.

Our argument follows the subsequent logic. First we introduce a weak refinement of the standard Bayes-Nash equilibrium which says, roughly speaking, that if some agent can unilaterally deviate to a strategy that does not reveal his deception ex post and do this in such a way that it does not affect the material outcome, while the original strategy of a Bayes-Nash equilibrium would reveal his dishonesty, then the mechanism designer can be confident that agents will not coordinate into this equilibrium.[4] After this we show that with our new solution concept the set of incentive compatible social choice functions coinsides exactly the set of Bayesian implementable social choice functions. Thus, as this result suggest, there is nothing wrong in relying on the revelation principle from an operational point of view. Incentive compatibility is a necessary condition for both implementation and

[2]Psychologists have observed that lying is cognitively more expensive than truth-telling (Greene and Paxton, 2009). It is possible that this is due to the social norm against lying since one has to be more careful when lying so that one does not get caught.

[3]St. Augustine wrote in 421: "To me, however, it seems certain that every lie is a sin ..." (Gneezy, 2005). Also, see Shariff and Norenzayan (2007) for an experiment where people lie less when primed for God concepts.

[4]That is, this Bayes-Nash equilibrium is not going to become focal (Schelling, 1960).

realization, while in the presence of social norm against lying it is sufficient as well, and therefore it is completely legitimate to restrict the search for a best social choice function into the set of incentive compatible direct mechanisms. Although, we may need a more complex construction to actually implement the social choice functions if we want to avoid all bad equilibria.

The rest of this paper is organized in the following way. In Section 2 we formulate the general Bayesian mechanism design problem and give an exact statement of the revelation principle. In Section 3 we make a little detour on the existing literature on Bayesian implementation before we give the result which implies that there is nothing wrong in relying on the revelation principle. The actual proof is relegated to the Appendix. Section 4 presents some further connections with the literature, while Section 5 concludes the paper with a brief discussion.

2 A general mechanism design setting with a numéraire

There is a finite group of agents that interact to make a joint decision. We denote the set of *agents* by $N = \{1, 2, \ldots, n\}$, with a generic element represented by i, j or k, and the set of *decision* by D, with a generic element represented by d or d'. All agents hold private information and the information of agent i is represented by a *type* θ_i that lies in a finite set Θ_i. A *state* is any profile of types $\theta = (\theta_1, \ldots, \theta_n) \in \Theta = \underset{i \in N}{\times} \Theta_i$. Let $\Theta_{-i} = \underset{j \neq i}{\times} \Theta_j$ and $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ as usual. We assume that there is a common prior belief $q(\cdot)$ over the set of states Θ. At the interim stage, after the type θ_i of agent i has been realized, beliefs are updated using the *Bayes' rule*

$$q\left(\theta_{-i} \mid \theta_i\right) = \frac{q(\theta_{-i}, \theta_i)}{\sum\limits_{\theta'_{-i} \in \Theta_{-i}} q(\theta'_{-i}, \theta_i)}.$$

For notational simplicity we assume that $Supp(\Theta) = \{\theta \in \Theta \mid q(\theta) > 0\} = \Theta$. Otherwise we would need to qualify everything by saying that it holds in the support of Θ.[5]

A *decision rule* (or an *allocation rule*) is a mapping $p : \Theta \to D$ that selects a decision $p(\theta) \in D$ as a function of the state $\theta \in \Theta$. In order to provide incentives it is possible for the mechanism designer to tax or subsidy agents. This is represented by a *transfer function* $t : \Theta \to \mathbb{R}^n$, where $t_i(\theta)$ is the *expected payment* that agent i receives or makes (if negative) when the state is $\theta \in \Theta$. A *social choice function* (SCF) $f : \Theta \to D \times \mathbb{R}^n$ is any mapping $f = (p, t)$ that is defined by giving a decision rule p together with a transfer function t. Each agent has a preference over decisions and money representable by a utility function $u_i : D \times \mathbb{R}^n \times \Theta \to \mathbb{R}$ that we assume to be quasi-linear in money. Thus, we can write $u_i(d, t; \theta) = v_i(d, \theta) + t_i$ and $u_i(d, t; \theta) > u_i(d', t'; \theta)$ indicates that agent i prefers (d, t) to (d', t') when the state is $\theta \in \Theta$.[6]

To complete the model we need to define how agents select when they face an uncertain prospect. To this end, let F be the set of all possible SCFs. A generic element of this set will be denoted by f, g or h. The utility function u_i together with the prior belief $q(\cdot)$ determines an (interim) expected utility of SCF $f = (p, t) \in F$ for agent i with type $\theta_i \in \Theta_i$ as

$$U_i\big(f; \theta_i\big) = \sum_{\theta_{-i} \in \Theta_{-i}} \Big[v_i\Big(p(\theta_i, \theta_{-i}); (\theta_{-i}, \theta_i)\Big) - t_i(\theta_i, \theta_{-i}) \Big] q(\theta_{-i} \mid \theta_i).$$

When $U_i\big(f; \theta_i\big) \geq U_i\big(g; \theta_i\big)$, we say that agent i with type θ_i weakly prefers f to g. Henceforth we denote this preference relation by $R^i(\theta_i)$, with a strict part $P^i(\theta_i)$ and indifference part $I^i(\theta_i)$ respectively. We call $E = \big(N, F, \Theta, q(\cdot), \{U_i\}\big)$ an *environment*.

A *mechanism* is a pair $\Gamma = (M, \mu)$, where $M = M_1 \times \cdots \times M_n$ is the *message space* and $\mu : M \to D \times \mathbb{R}^n$ is the *outcome function*. Thus, for any profile

[5] See Jackson (1991) for a definition of incentive compatibility in the case $Supp(\Theta) \neq \Theta$.

[6] The traditional *economic environment* -assumption (e.g. Jackson, 1991, Palfrey and Srivastava, 1993) is not sufficient for our mechanism to work properly. We elaborate more on this issue in the Appendix.

of messages $m = (m_1, \ldots, m_n) \in M$, $\mu(m) = (\mu_0(m), \mu_1(m), \ldots, \mu_n(m))$ is the resulting decision $\mu_0(m) \in D$ together with the expected transfers $\mu_1(m), \ldots, \mu_n(m) \in \mathbb{R}$. A *strategy* of agent i is a function $\sigma_i : \Theta_i \to M_i$. We write Σ_i for the set of all strategies of agent i and $\Sigma = \times_{i=1}^{n} \Sigma_i$ for the set of all strategy profiles. A strategy profile $\sigma = (\sigma_1, \ldots, \sigma_n) \in \Sigma$ is a *Bayes-Nash equilibrium* of Γ if, and only if, $\mu(\sigma_i, \sigma_{-i}) R^i(\theta_i) \mu(m_i, \sigma_{-i})$ for all $i \in N$, $\theta_i \in \Theta_i$ and $m_i \in M_i$.[7] Let us denote the set of all Bayes-Nash equilibria in Γ by $BNE(\Gamma)$.

We say that mechanism Γ *realizes the SCF* $f \in F$ *in Bayes-Nash equilibrium* if there exists $\sigma \in BNE(\Gamma)$, such that $\mu(\sigma(\theta)) = f(\theta)$ for all $\theta \in \Theta$. Sometimes, however, the mechanism designer may have a stronger objective in mind. We say that mechanism Γ *implements the SCF* $f \in F$ *in Bayes-Nash equilibrium* if $BNE(\Gamma) \neq \emptyset$, and for each $\sigma \in BNE(\Gamma)$, we have $\mu(\sigma(\theta)) = f(\theta)$ for all $\theta \in \Theta$. As usual, we say that f is *Bayesian implementable* if there exists a mechanism Γ that implements f in Bayes-Nash equilibrium .[8]

A *direct mechanism* is a mechanism $\Gamma^f = (\Theta, f)$, where the message space is the set of states $\Theta = \Theta_1 \times \cdots \times \Theta_n$ and the outcome function is a SCF f. In other words, a direct mechanism simply ask agents to announce their type and then select whatever outcome is recommended by the SCF. From now on we denote the truthful strategy profile by $\widehat{\sigma} = (\widehat{\sigma}_1, \ldots, \widehat{\sigma}_n)$, so that $\widehat{\sigma}_i(\theta_i) = \theta_i$ for all $i \in N$ and all $\theta_i \in \Theta_i$. A SCF f is called *incentive compatible* (IC) if the truthful strategy profile $\widehat{\sigma}$ is a Bayes-Nash equilibrium of the associated direct mechanism Γ^f, that is if

$$f(\theta_i, \widehat{\sigma}_{-i}) R^i(\theta_i) f(\theta_i', \widehat{\sigma}_{-i}) \quad \text{for all } i \in N \text{ and all } \theta_i, \theta_i' \in \Theta_i.$$

This says that when all agents except one are telling the truth, the remaining agent does not have an incentive to lie either. We are finally ready to state

[7] See Harsanyi (1967-68) for an in depth analysis of this equilibrium concept.

[8] Mookherjee and Reichelstein (1990), Jackson (1991), Palfrey and Srivastava (1993) and Duggan (1995) provide characterizations of Bayesian implementable SCFs and more generally of social choice correspondences.

the celebrated revelation principle.[9]

The Revelation Principle for Bayes-Nash Equilibrium: If mechanism $\Gamma = (M, \mu)$ realizes (or implements) the SCF $f \in F$ in Bayes-Nash equilibrium, then the truthful strategy profile $\widehat{\sigma}$ must be an equilibrium of the associated direct mechanism Γ^f. In other words, social choice function f must be IC. ■

This result is famous for its ability to make hard problems tractable. It tells us that instead of looking at all possible mechanisms we can restrict attention to truthful strategies of incentive compatible direct mechanisms. Unfortunately, it has been shown many times by many different authors that an incentive compatible direct mechanism can have other equilibria besides the truthful one. This means that sometimes there exists another Bayes-Nash equilibrium in which multiple agents are lying simultaneously, and it may not even be possible to implement a SCF that can be realized using an incentive compatible direct mechanism (Jackson, 1991; Duggan, 1995; Bergin, 1995).[10] That is to say, any mechanism, direct or indirect, will always have another equilibrium that does not coincide with the SCF. Therefore, revelation principle relies heavily on the truthful strategy profile being somehow focal, or the mechanism designer at least being able to make it so.[11]

3 Bayesian implementation with social norm against lying

Although the multiple equilibrium problem is unresolvable in a strict sense, it is still possible that the problem is not significant since all equilibria may not be equally plausible. After all, this is exactly what the alleged focality of truthful strategies suggests. Recently, in Korpela (2014), it was shown that

[9] See Dasgupta et al. (1979) for a proof.

[10] Since IC is not sufficient for implementation in the standard sense.

[11] There are nice stories in Myerson (2009) which suggest that nearly any equilibrium can become focal.

in any standard resource allocation problem where all agents are partially honest, which means, roughly speaking, that an agent does not lie unless it affects his material payoff, the set of incentive compatible SCFs coincides exactly with the set of implementable SCFs.[12] In other word, if the focality of truth-telling means that all agents are partially honest, then the revelation principle works just fine as a practical tool.

Unfortunately, although agents are partially honest in some environments, this is hardly an assumption that can claim general validity. To make a point, suppose that 10 agents are deciding whether to undertake a joint project. The decision rule is binary with two possible outcomes - start the project or not. Let us furthermore assume that the decision is made via majority voting and that all agents are of two possible types, either poor or rich, with equal probability. Poor agents prefer not to undertake the project and rich agents prefer to undertake it. Now suppose, for the sake of argument, that for whatever reason everyone expects the other 9 agents to claim that they are rich irrespective of their true type. Would an agent in this situation claim that he is poor although he has a firm belief that this will not affect the outcome? Possibly, but there is no reason to regard this as certain. Seeing no way to affect the outcome agents might very well want to pose as rich. The only thing that we can be sure of is that once an agent decides to lie he will not want to get caught.[13]

In line with this intuition, we acknowledge that people lie but prefer to do it in a consistent way rather than in an inconsistent way. To define the distinction suppose that the message space of agent i is $M_i = \Theta_i \times Q_i$. Moreover, let us divide the strategy $\sigma_i : \Theta_i \to M_i$ of agent i in two parts $\sigma_i = (\alpha_i, s_i)$, where $\alpha_i : \Theta_i \to \Theta_i$ is the *deception* and $s_i : \Theta_i \to Q_i$ is the *auxiliary strategy*. There are two types of deceptions, those in which agent gets caught lying ex post, and those in which agent does not get caught lying

[12] See also Dutta and Sen (2012), Lombardi and Yoshihara (2011), Kartik, Holden and Tercieux (2014) and Doghmi and Ziad (2013). All these papers, however, study the concept mainly under complete information.

[13] Except admittedly in some rare cases.

or did not lie in the first place. Formally, if agent i with type θ_i lies that his type is θ_i' instead, that is $\alpha_i(\theta_i) = \theta_i'$, and at the same time with type θ_i' lies that his type is θ_i'' instead, that is $\alpha_i(\theta_i') = \theta_i''$, then the agent will get caught lying ex post when the true type is θ_i since others do not expect to see θ_i' played under θ_i'. On the other hand, if agent i does not lie when his type is θ_i', that is $\alpha_i(\theta_i') = \theta_i'$, then he does not get caught lying under θ_i as the true type could be θ_i' instead of θ_i. That is, others cannot tell whether he is lying or not, or they cannot at least accuse him of such behavior.

Mathematically the difference is whether the deception is an idempotent function or not.

Definition 1. The deception α_i is *idempotent* if $\alpha_i \circ \alpha_i \equiv \alpha_i$. In all other cases it is called *non-idempotent* and there will then exist some type $\theta_i \in \Theta_i$ such that $\alpha_i(\alpha_i(\theta_i)) \neq \alpha_i(\theta_i)$. An idempotent deception is called *consistent*, in line with the lying interpretation, and a non-idempotent deception is called *inconsistent*. □

In other words, if the deception of agent i is inconsistent, then some type of this agent mimics another type that would not be truthful either and therefore is caught lying ex post. With this definition in mind we can express the idea that some Bayes-Nash equilibria will not be played in the presence of a social norm against lying. We say that $\sigma_i = (\alpha_i, s_i)$ is consistent if α_i is, and that $\sigma = (\sigma_1, \ldots, \sigma_n)$ is consistent if σ_i is consistent for all $i \in N$.

Definition 2. Let $\Gamma = (M, g)$ be a mechanism, with $M_i = \Theta_i \times Q_i$ for all $i \in N$. Here Q_i is the part of message space that does not have any special meaning attached to it. We say that $\sigma \in BNE(\Gamma)$ is *not* an acceptable Bayes-Nash equilibrium when there is a social norm against lying if the following two conditions hold:

(1) There exists an agent $i \in N$, such that $g\big(\sigma_i, \sigma_{-i}\big) I^i(\theta_i) g\big(\sigma_i', \sigma_{-i}\big)$ for all $\theta_i \in \Theta_i$, and

(2) σ_i' is consistent, while σ_i is not.

□

One way to think about this definition is as a weak refinement of the standard Bayes-Nash equilibrium.[14] It says that agents have a strict aversion towards inconsistent deceptions provided their expected payoffs are not affected. In the language of Schelling (1960), if both item (1) and (2) in Definition 2 are satisfied, then the mechanism designer can be confident that σ is not going to turn out as focal. Notice that in this definition we implicitly assume that agents cannot break the norm against lying by themselves, rather the norm has its origin outside the mechanism.[15] If the agents cannot affect the norm, then it is better to retain a honest appearance. Thus, it is reasonable to assume that although agents do not care about whether they lie or not *per se*, they would still prefer to lie in a consistent way. If we return to the example of agents voting on a joint project, Definition 2 does not claim that these agent would never lie when this does not affect the material outcome, rather it simply states that, yes, agents can be expected to lie, but they prefer to do it in a way consistent with the contingency that they find themselves in.[16]

To study implementation under this refined equilibrium we need to define what is meant by Bayesian implementation with a social norm against lying. Let us denote the set of all Bayes-Nash equilibria which are acceptable in the sense of Definition 2 by $BNE^{+}(\Gamma)$. We say that mechanism Γ *realizes the SCF* $f \in F$ *in Bayes-Nash equilibrium with a social norm against lying* if there exists at least one $\sigma \in BNE^{+}(\Gamma)$, such that $\mu(\sigma(\theta)) = f(\theta)$ for all $\theta \in \Theta$. Similarly, we say that mechanism Γ *implements the SCF* $f \in F$ *in Bayes-Nash equilibrium with a social norm against lying* if for every equilibrium $\sigma \in BNE^{+}(\Gamma)$, we have that $\mu(\sigma(\theta)) = f(\theta)$ for all $\theta \in \Theta$, and $BNE^{+}(\Gamma) \neq \emptyset$. Thus, the idea of realization and implementation generalize in a straightforward way.

It is important to keep in mind that IC is a necessary condition for realiza-

[14] The fact that our refinement is as weak as possible will reinforce the eventual conclusion.

[15] In this sense N cannot be the entire society, rather it has to be a small sub-population.

[16] See Diekmann et al. (2011) however.

tion and implementation also under our new solution concept, and therefore revelation principle continues to hold. The following Theorem and Corollary will wrap up the discussion. Although, our Theorem is more like an argument, or observation, or perhaps just a note, than anything else. However, it do suggest that the so called multiple equilibrium problem may not be at all that fatal.

THEOREM. If an SCF f is incentive compatible, then it is implementable in Bayes-Nash equilibrium with a social norm against lying.

Proof. See the Appendix. ■

To summarise, suppose that all other things equal a dishonest person would always prefer to lie in a consistent way. Under this assumption, if a mechanism can realize a SCF, then it must be IC. On the other hand, by the above Theorem we know that any incentive compatible SCF can be implemented in Bayes-Nash equilibrium with a social norm against lying. Thus, we can make the following conclusion.

COROLLARY. In a methodological sense there is nothing wrong in relying on the revelation principle. Excluding implausible equilibria, the set of incentive compatible SCFs coincides exactly with the set of implementable SCFs. It is just that a direct revelation mechanism may not be able to implement the SCF and a more complex construction is needed. ■

4 Further connections with the literature

We have studied revelation principle in the simplest case of one principal and multiple agents. Saran (2011) has recently shown that in this case revelation principle is even robust to deviations from the rational framework. However, if there are many competing principals, then the revelation principle itself may no longer hold (Peters and Epstein, 1999; Peters, 2001).

The burgeoning literature on behavioral economics is full of results that have potential applications in mechanism design (see e.g. DellaVigna, 2009;

Diamond and Vartiainen, 2007). Two that are important here are those obtained in the literature on lying costs (Kartik, 2009b; Abeler et al., 2014) and contagiousness of norm violations (Diekmann et al., 2011; Houser et al., 2012). The literature on lying costs suggest that IC is not always necessary for implementation because agents may not care only about the material payoff.[17] The literature on contagiousness of norm violations, on the other hand, has a direct relevance to what we are doing here because one of our key assumptions is that agents cannot break the norm against lying by themselves.

Another obvious possibility would be to consider models of bounded rationality (see Rubinstein, 1998). One notable paper in this respect is de Clippel et al. (2014) who get similar result as we do here by assuming a bounded depth of reasoning.[18] Our result, however, derives from exactly the opposite direction. We assume that agents are rational in the standard sense, and in addition, take the prevailing social norms against lying into account when making decisions. Thus, agents in our model could be characterized as superrational rather than boundedly rational or behavioral.

The paper that is closest to ours in spirit is Matsushima (1993). He shows that if a condition called *no consistent deceptions* (NCD) is satisfied, then IC is a full characterization of Bayesian implementable SCFs in any economic environment.[19] In contrast to our paper, and despite its name, NCD is in fact a restriction on the information structure $q(\cdot)$ rather than on the deceptions. Furthermore, it is not entirely obvious why some information structures should be ruled out *a priori*. However, the idea that all deceptions are not alike is clearly present in this paper.

[17] There are many other reasons why IC is not always necessary for implementation, see Glazer and Rubinstein (2013) and Kartik et al. (2014) for example.

[18] See de Clippel (2014) for a great introduction to behavioral implementation.

[19] Matsushima (1993) does not use the word *consistent* in the same way as we do here.

5 Conclusion

We have argued that IC is a full characterization of Bayesian implementable SCFs in any society that has a weak norm against lying. This result holds also in the case of two players, and moreover, it implies that there is nothing operationally wrong in relying on the revelation principle. However, it is not directly applicable to the myriad of results obtained in the literature on auction design (Krishna, 2002; Milgrom, 2004) since we have assumed a finite type space. In fact, our mechanism would not work as such if the type space would be infinite. In this regard it can only be considered as suggestive.

One nice feature of the mechanism that we use is that it has exactly one equilibrium. Thus, implementation cannot fail simply because agents try to coordinate in different equilibria. This is reassuring since the possibility that agents do not coordinate into the same equilibrium grows dramatically as the number of agents increase. However, we are using integer games, which is another reason why the result can be considered only as suggestive.

APPENDIX: Proof of the Theorem

Suppose that f is IC. We construct a mechanism that implements f in Bayes-Nash equilibrium with a social norm against lying. Let D_i be the set of all deceptions of agent i and assume that agents are placed on a circle in such a way that 1 is between 2 and n. Thus, the predecessor of agent 1 is agent n and we denote $0 = n$ accordingly. The message space of agent i is $M_i = \Theta_i \times \left\{\Theta_{i-1} \cup \{T\}\right\} \times D_i \times \{0,1,2\} \times \mathbb{N}$, with a typical message denoted by $m^i = (\theta_i, \theta_{i-1}, \alpha_i, k_i, n_i)$, while the outcome function μ is defined via the following rules:[20]

(**1**) If $m_4^i = 0$ for all $i \in N \setminus \{j\}$, and $m_4^j = 0$ or $\left[m_4^j = 1 \text{ and } m_2^j = T\right]$, then

$$\mu(m) = f\Big(m_1^1, \ldots, m_1^i, \ldots, m_1^n\Big).$$

(**2**) If there exists $j \in N$, such that $m_4^i = 0$ for all $i \in N \setminus \{j\}$, but $m_4^j = 1$ and $m_2^j \neq T$, then the outcome is determined as in Rule (1) except that

$$\mu(m)_j = \begin{cases} f\Big(m_1^1, \ldots, m_1^i, \ldots, m_1^n\Big)_j + \epsilon & \text{if } m_2^j = \theta_{j-1} \neq m_1^{j-1} \\ f\Big(m_1^1, \ldots, m_1^i, \ldots, m_1^n\Big)_j - A & \text{if } m_2^j = \theta_{j-1} = m_1^{j-1}. \end{cases}$$

(here A must be large enough so that agent j will never want to deviate from Rule (1) if there is a possibility to incur the loss.[21])

(**3**) If there exists $j \in N$, such that $m_4^i = 0$ for all $i \in N \setminus \{j\}$, but $m_4^j = 2$ and $m_3^j = \alpha_j$, then

$$\mu(m) = f\Big(m_1^1, \ldots, \alpha_j(m_1^j), \ldots, m_1^n\Big).$$

(**4**) In all other cases, denote $k = \max\{i \mid m_5^i \geq m_5^j \text{ for all } j = 1, \ldots, n\}$, and let the outcome be otherwise as in Rule (1) except that

$$\mu(m)_k = f\Big(m_1^1, \ldots, m_1^n\Big)_k + \epsilon.$$

[20] Remember that the outcome $\mu(m)$ is an $(n+1)$ -tuple where the first component (the decision) is denoted by $\mu(m)_0$.

[21] Denote $q' = \min\{q(\theta) \mid \theta \in \Theta\}$. One possibility is to select A and ϵ in such a way that $q'A > \epsilon$.

REMARK 1. In this mechanism agent i can use the set $\Theta_{i-1} \cup T$ to indicate that he believes agent $i-1$ is truthful in general (T) or lying under some type (Θ_{i-1}). This does not mean that agent i has to know the type of agent $i-1$. On the contrary, agent i only needs to have a belief that certain types are never part of any message sent by agent $i-1$. Furthermore, we can use two bijective mappings, $\psi : \Theta_{i-1} \cup T \to Q$ and $\varphi : D_i \to Q'$, to code the sets $\Theta_{i-1} \cup T$ and D_i into something that does not have any special meaning attached to it. In other words, we use $\left\{\Theta_{i-1} \cup T\right\} \times D_i \times \{0,1,2\} \times \mathbb{N}$ only as an accounting device to simplify the analysis. ‖

REMARK 2. Deceptions can be divided into two disjoint sets - permutations and the rest. If agent i is using any consistent deception other than the identity mapping under Rule (1), then it is profitable for agent $i+1$ to deviate into Rule (2) since the deception cannot be a permutation and therefore some type of agent i is never part of any message. On the other hand, agents do not want to use inconsistent deception under Rule (1) either since it is possible to obtain exactly the same outcome with a consistent deception under Rule (3). Social norm against lying guarantees that all agents prefer a consistent deception to an inconsistent one. ‖

REMARK 3. Our mechanism does not work under the commonly used *economic environment* -assumption. Roughly speaking, an economic environment means that the same decision cannot be the best alternative of any $n-1$ agents at the same time. This assumption is satisfied, for example, by any standard resource allocation problem where the best alternative of every agent is to get all resources to himself. However, it does not guarantee that we can find the outcomes that are needed in Rule (2). ‖

Next we prove that this mechanism implements f. The proof proceeds in the following way. First we show that a strategy profile $\sigma = (\sigma_1, \ldots, \sigma_n) = ((\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n))$ can be a Bayes-Nash equilibrium with a social norm against lying if, and only if, the outcome is always selected using Rule (1). After this we show that at any equilibrium under Rule (1) all deceptions α_i must be identity mappings and therefore the outcome coincides with the SCF

f. This is done by verifying two things: (a) if some agent, say agent i, would be using a consistent deception, then agent $i+1$ would rather deviate to Rule (2), and (b) it is not possible that agents are using inconsistent deception either since then they would rather deviate to Rule (3) themselves.

Proof. Suppose that $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a Bayes-Nash equilibrium with a social norm against lying. The outcome is never selected using Rules (2), (3) or (4). First of all, at any state the outcome cannot be selected using Rule (4) since some agent could improve his position by increasing the fifth component and leaving everything else intact. This would guarantee that the outcome is exactly as before except that it would be better under Rule (4). Furthermore, there is always some agent who can deviate from Rules (2) and (3) to Rule (4) and improve his position in so doing. Therefore, all equilibria of this mechanism (if any) must select the outcome using Rule (1) at all possible states.

The fact that f is IC guarantees that if the outcome is always selected using Rule (1) and all agents are truthful at all states, then σ is a Bayes-Nash equilibrium with a social norm against lying. Thus, there is at least one good equilibrium since in this case the outcome coincides with f by definition. Moreover, the mechanism does not have any other equilibria besides this one. To argue for this let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be any profile of deception that can be used as a part of σ. Suppose that agent j is not truthful, so that α_j is not an identity mapping, and that the outcome is always selected using Rule (1). There are two possibilities: (a) α_j is a consistent deception or (b) α_j is an inconsistent deception. In the first case agent $j+1$ could deviate profitably to Rule (2) since there must be some type, say $\theta_j \in \Theta_j$, that does not belong to the range of α_j.[22] This means that agent $j+1$ could announce θ_j as the second component and 1 as the fourth component of his strategy and deviate to Rule (2) without any fear of suffering the penalty A. The outcome would be exactly as before except that the transfer would be higher. In the second case it is agent j himself that wants to deviate. This is because he can start telling the truth and deviate to Rule (3) by adjusting

[22] Here we need the assumption that Θ_j is finite.

the third and fourth component of his strategy in such a way that he gets exactly the same outcome as before. He prefers to do this because of the social norm against lying. This completes the proof. ■

References

Abeler, J., Becker, A., Falk, A. (2014): "Representative Evidence on Lying Costs". Forthcoming in *Journal of Public Economics*.

Amir, O., Ariely, D., Mazar, N. (2008): "The dishonesty of honest people: A theory of self-concept maintenance". *Journal of Marketing Research* **45**(6): 633-644.

Aumann, R. (1974): "Subjectivity and correlation in randomized strategies". *Journal of Mathematical Economics* **1**: 67-96.

Bergin, J. (1995): "On some recent results in incomplete information implementation". *Canadian Journal of Economics* **28**: 108-138.

Dasgupta, P., Hammond, P., Maskin, E. (1979): "The implementation of social choice rules: Some general results on incentive compatibility". *Review of Economic Studies* **46**(2): 185-216.

de Clippel, G. (2014): "Behavioral Implementation". Forthcoming in *American Economic Review*.

de Clippel, G., Saran, R., Serrano, R. (2014): Mechanism Design with Bounded Depth of Reasoning and Small Modeling Mistakes. Working Paper, *mimeo*.

Diekmann, A., Przepiorka, W., Rauhut, H. (2011): Lifting the veil of ignorance: An experiment on the contagiousness of norm violations. CESS Discussion Paper, *mimeo*.

Doghmi, A., Ziad, A. (2013): "On partially honest Nash implementation in private good economies with restricted domains: A sufficient condition". *The B. E. Journal of Theoretical Economics* **13**: pages 14.

Duggan, J. (1995): Bayesian implementation. Dissertation, California Institute of Technology, *mimeo*.

Dutta, B., Sen, A. (2012): "Nash implementation with partially honest individuals". *Games and Economic Behavior* **74**: 154-169.

Feldman, A., Serrano, R. (2006): Bayesian implementation. Chapter 16 in *Welfare Economics and Social Choice Theory* 2^{nd} *Edition*, Springer Science + Business Media, USA.

Fischbacher, U., Föllmi-Heusi, F. (2013): "Lies in disguise - An

experimental study on cheating". *Journal of the European Economic Association* **11**(3): 525-547.

Gibbard, A. (1973): "Manipulation of voting schemes: A general result". *Econometrica* **41**(4): 587-601.

Glazer, J., Rubinstein, A. (2013): Complex questionnaires. Working Paper, *mimeo.*

Gneezy, U. (2005): "Deception: The role of consequences". *American Economic Review* **95**(1): 385-394.

Greene, J., Paxton, J. (2009): "Patterns of neural activity associated with honest and dishonest moral decisions". *Proceedings of the National Academy of Science of the United States of America* **106**(30): 12506-12511.

Houser, D., Vetter, S., Winter, J. K. (2012): "Fairness and cheating". *European Economic Review* **56**: 1645-1655.

Houser, D., Hao, L. (2011): Honest lies. Interdisciplinary Center for Economic Science, George Mason University, Discussion Paper, *mimeo.*

Harris, M., Townsend, R. (1981): "Resource allocation under asymmetric information". *Econometrica* **49**(1): 33-64.

Harsanyi, J. (1967-68): "Games with Incomplete Information Played by 'Bayesian' Players". *Management Science* **14**: 159-189, 320-334, 486-502.

Holmström, B. (1977): "On incentives and control in organizations". Dissertation, Stanford University.

Jackson, M. O. (1991): "Bayesian implementation". *Econometrica* **59**: 461-477.

Kartik, N., Holden, R., Tercieux, O. (2014): "Simple mechanisms and preference for honesty". *Games and Economic Behavior* **83**: 284-290.

Kartik, N., Hurkens, S. (2009a): "Would I lie to you? On social preferences and lying aversion". *Experimental Economics* **12**: 180-192.

Kartik, N. (2009b): "Strategic communication with lying costs". *Review of Economic Studies* **76**(4): 1359-1395.

Korpela, V. (2014): "Bayesian implementation with partially honest individuals". Forthcoming in *Social Choice and Welfare.*

Krishna, V. (2002): *Auction Theory*. Academic Press, USA.

Lebrun, B. (2006): "Uniqueness of the equilibrium in first-price auctions". *Games and Economic Behavior* **55**: 131-151.

Lombardi, M., Yoshihara, N. (2011): Partially-honest Nash implementation: Characterization results. Faculty of Business, Economics and Law, University of Surrey, *mimeo*.

Maskin, E., Riley, J. (2003): "Uniqueness of equilibrium in sealed high-bid auctions". *Games and Economic Behavior* **45**: 395-409.

Matsushima, H. (2008a): "Behavioral aspects of implementation theory". *Economics Letters* **100**: 161-164.

Matsushima, H. (2008b): "Role of honesty in full implementation". *Journal of Economic Theory* **139**: 353-359.

Matsushima, H. (1993): "Bayesian implementation with side payments". *Journal of Economic Theory* **59**: 107-121.

Milgrom, P. (2004): *Putting Auction Theory to Work*. Cambridge University Press, UK.

Mookherjee, D., Reichelstein, S. (1990): "Implementation via augmented revelation mechanisms". *Review of Economic Studies* **57**: 453-475.

Myerson, R. (2009): "Learning from Schelling's Strategy of Conflict". *Journal of Economic Literature* **47**(4): 1109-25.

Myerson, R. (2008): "Perspectives on mechanism design in economic theory". *American Economic Review* **98**(3): 586-603.

Myerson, R. (1981): "Optimal auction design". *Mathematics of Operations Research* **6**: 58-73.

Myerson, R. (1979): "Incentive compatibility and the bargaining problem". *Econometrica* **47**(1): 61-74.

Peters, M., Epstein, L. (1999): "A Revelation Principle for Competing Mechanisms". *Journal of Economic Theory* **88**: 119-160.

Peters, M. (2001): "Common Agency and the Revelation Principle". *Econometrica* **69**(5): 1349-1372.

Palfrey, T., Srivastava, S. (1993): *Bayesian Implementation*. Harwood

Academic Publishers, Switzerland.

Palfrey, T. (1992): Implementation in Bayesian Equilibrium - The Multiple Equilibrium Problem in Mechanism Design. In *Advances in Economic Theory: Sixth World Congress*, Vol. I, edited by J.-J. Laffont. Cambridge, UK, Cambridge University Press.

Radner, R, Ray, D. (2003): "Robert W. Rosenthal". *Journal of Economic Theory* **112**: 365-368.

Rosenthal, R. (1978): "Arbitration of two-party dispute under uncertainty". *Review of Economic Studies* **45**(3): 595-604.

Rubinstein, A. (1998): *Modeling bounded rationality*. MIT Press, USA.

Saran, R. (2011): "Menu-dependent preferences and revelation principle". *Journal of Economic Theory* **146**: 1712-1720.

Schelling, T. (1960): *The Strategy of Conflict*. Harvard University Press, Cambridge.

Shariff, A. F., Norenzayan, A. (2007): "God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game". *Psychological Science* **18**(9): 803-809.

The **Aboa Centre for Economics (ACE)** is a joint initiative of the economics departments of the Turku School of Economics at the University of Turku and the School of Business and Economics at Åbo Akademi University. ACE was founded in 1998. The aim of the Centre is to coordinate research and education related to economics.

Contact information: Aboa Centre for Economics, Department of Economics, Rehtorinpellonkatu 3, FI-20500 Turku, Finland.

www.ace-economics.fi

ISSN 1796-3133