Journal of Mathematical Economics 78 (2018) 27–34

Contents lists available at ScienceDirect

Journal of Mathematical Economics

journal homepage: www.elsevier.com/locate/jmateco

# Strong implementation with partially honest individuals

Foivos Savva

*Adam Smith Business School, University of Glasgow, West Quadrangle, Gilbert Scott Building, Glasgow, G12 8QQ, United Kingdom*

## HIGHLIGHTS

- We provide sufficient conditions for strong implementation with partially honesty.
- No Maskin monotonicity type condition is used.
- Man-optimal stable solution in pure matching problems is strongly implementable.
- Nash bargaining solution is strongly implementable.

## ARTICLE INFO

*Article history:*
Received 12 August 2017
Received in revised form 2 July 2018
Accepted 5 July 2018
Available online 17 July 2018

*Keywords:*
Strong implementation
Partial honesty
Tie-breaking rule

## ABSTRACT

In this paper we provide sufficient conditions for a social choice rule to be implementable in strong Nash equilibrium in the presence of partially honest agents, that is, agents who break ties in favour of a truthful message when they face indifference between outcomes. In this way, we achieve a relaxation in the condition of Korpela (2013), namely the *Axiom of Sufficient Reason*. Our new condition, *Weak Pareto Dominance*, is shown to be sufficient along with *Weak Pareto Optimality* and *Universally Worst Alternative*. We finally provide applications of our result in pure matching and bargaining environments.



## 1. Introduction

Implementation theory studies the relationship between social goals and institutions.[1] Specifically, it aims to examine the effect of institutional design to the attainment of socially desirable outcomes. For example, suppose that a group of people have agreed on the desirable social outcomes as a function of their preferences. How can they make sure that they can indeed obtain those outcomes, when some or all of them may potentially benefit by misrepresenting their preferences? They thus have to rely on designing an institution (in other words, mechanism or game form) through which they will interact, that will ensure the optimality of the outcomes reached through this interaction. More formally, for any collective choice rule that assigns some socially optimal outcomes as a function of individual preferences, implementation is achieved when, for any profile of preferences, the set of socially optimal outcomes coincides with the set of outcomes attained in the equilibrium of the game induced by the mechanism.

While most of the classic literature on the subject relies on the assumption that agents have a purely *consequentialist* nature, that is, they only care about the final outcomes, the strand of behavioural implementation theory typically assumes that agents may also have *procedural* concerns. One recent subfield in particular, takes into account the fact that agents may have an intrinsic preference for honesty. This weak honesty motive is usually modelled in the following manner: Suppose that an agent is indifferent between two outcomes. Then she will strictly prefer to obtain an outcome with a truthful message rather than with an untruthful one. This type of rationale is typically referred to as *partial honesty* or *minimal honesty* and can be supported by the experimental findings of Hurkens and Kartik (2009) for example, who show that subjects either are always honest, or tend to lie only when they gain by doing so. Despite being rather weak, partial honesty is shown to bear a significant positive effect for the set of implementable rules and limitations imposed by *Maskin-monotonicity*[2] in particular. In their seminal paper, Dutta and Sen (2012) show that in the presence of at least one partially honest agent in the society, *Maskin-monotonicity* is no longer a necessary condition for Nash implementation and *No Veto Power* alone becomes sufficient for three or more agents.

*E-mail address:* f.savva.2@research.gla.ac.uk.

[1] For a comprehensive survey of the main results in the literature of implementation theory see Jackson (2001).

[2] Maskin (1999) in his seminal paper identified a condition now known as *Maskin-monotonicity* as necessary and almost sufficient for Nash implementation. It roughly says that if an optimal outcome at some state does not fall in even one person's ranking when switching to another state, then it should still be selected as optimal. A formal definition will be given later.

https://doi.org/10.1016/j.jmateco.2018.07.002

Overall, the results on Nash implementation with partial honesty have been positive. An important question that remains unanswered though is whether these possibilities can be extended to other, possibly stronger, equilibrium concepts. For example, in many situations, the social planner cannot exclude the possibility of pre play communication between the agents and thus the mechanism may be vulnerable to group deviations. In such settings the natural solution concept to use is strong Nash equilibrium[3] *à la* (Aumann, 1960), that is robust to deviations by any possible coalition of agents.

The current paper identifies sufficient conditions for strong implementation when all agents are partially honest. Instead of a full characterization, we chose to follow the work of Korpela (2013) in providing simple sufficient conditions that have a more intuitive appeal and are generally easier to check in applications. First, we identify sufficient conditions for strong implementation when all agents are partially honest and prove their sufficiency. Specifically, we show that if a social choice rule satisfies *Weak Pareto Optimality* (WPO), *Universally Worst Alternative* (UWA) and *Weak Pareto Dominance* (WPD), then it can be implemented in strong equilibrium. In this way we achieve a relaxation in the condition of Korpela (2013), namely the *Axiom of Sufficient Reason* (ASR). Our new condition, WPD roughly requires the following to be true: if an outcome $a$ is optimal at some state, and if there exists another outcome $b$, such that all agents weakly prefer $b$ to $a$ with at least one agent being indifferent between them, then $b$ must be optimal as well. WPD is implied by ASR, therefore our condition is weaker. Next, we provide two applications of our results, in bargaining and pure matching environments. More specifically, we show that the man-optimal (or woman-optimal) solution in a pure matching environment as well as the Nash bargaining solution in a cake-cutting environment are both strongly implementable, when agents are partially honest.

The remainder of the paper is organized as follows: In Section 2, we review the relevant literature. In Section 3, we present the basic implementation setting and formal definitions. In Section 4, we provide the definitions of our conditions, our main theorem and some additional results. Section 5 consists of our two applications. Finally, in Section 6 we conclude by discussing our results and providing some points for further research. The proof of our main theorem is in the Appendix.

## 2. Related literature

The problem of strong implementation has primarily been studied by Maskin (1979). Moulin and Peleg (1982) provide some results on the same issue with the use of effectivity functions. A complete characterization of strongly implementable social choice rules is due to Dutta and Sen (1991). Suh (1996) generalizes the latter result by allowing the planner to possibly exclude some coalition formation *ex ante*, so in this more general setting not all coalitions are feasible. If the planner though cannot obtain such information, the relevant implementation concept is double implementation in Nash and strong equilibrium. Suh (1997) provides general results in this case as well. While complete characterizations are of high theoretical significance, they can be hard to apply to more specific settings. This motivates the more recent work by Korpela (2013) to identify simple sufficient conditions for strong implementation.

On the issue of partial honesty in implementation, the pioneering work of Dutta and Sen (2012) shows that *No Veto Power* (NVP) alone becomes sufficient for Nash implementation in the presence of at least one partially honest agent.[4] Their results are generalized by Lombardi and Yoshihara (2017b), who provide a full characterization of Nash implementable rules in the presence of partial honesty, for both unanimous and non-unanimous social choice rules. In more applied settings, Kartik et al. (2014) focus on environments with economic interest and identify sufficient conditions for implementation in two rounds of iterative deletion of strictly dominated strategies by "simple" mechanisms, without utilizing the usual *canonical mechanisms*.[5] On restricted domains with private goods, Doghmi and Ziad (2013) provide more positive results for Nash implementation. In other solution concepts with complete information, Saporiti (2014) shows that with partial honesty strategy-proofness is necessary and sufficient for secure implementation, which essentially requires implementation in dominant strategies and Nash equilibrium. Hagiwara (2017) also shows that NVP is sufficient with at least one, and unanimity is sufficient with at least two partially honest agents for double implementation in Nash and undominated Nash equilibria. Finally, the limitations of partial honesty in Nash implementation are outlined in Lombardi and Yoshihara (2018) who explore under which conditions partially honest Nash implementation is equivalent to Nash implementation, and in Adachi (2017).

Partial honesty can yield positive results in incomplete information environments as well. For example, Matsushima (2008) shows that incentive compatibility is sufficient for implementation in strong iterative dominance and Korpela (2014) proves that incentive compatibility and NVP are sufficient for implementation in Bayes Nash equilibrium. Studies with alternative solution concepts include Ortner (2015), who provides more positive results with partial honesty in fault-tolerant Nash equilibrium[6] and stochastically stable equilibrium.

The issue of implementation with partial honesty nevertheless can be put in the broader context of implementation with motives, where it is typically assumed that agents may also give significance to motives as procedural concerns, apart from the final outcomes. Along this line of research, it is worth mentioning a concept related to partial honesty, namely that of "social responsibility". In Lombardi and Yoshihara (2017a), the effect of social responsibility is explored with regards to natural implementation.[7] Hagiwara et al. (2018) utilize a similar concept of social responsibility for strategy space reduction with an outcome mechanism for Nash implementation. In a different environment, Doğan (2017) shows that the unique socially optimal allocation of objects to agents can be Nash implemented, when at least three agents have a social responsibility motive. Some general results on motives as tie-breaking rules with regards to Nash implementation are in Kimya (2017). Other significant contributions to the literature of motives in implementation include Glazer and Rubinstein (1998), Corchón and Herrero (2004) and Bierbrauer and Netzer (2016).

[3] From now on we will use the terms strong equilibrium and strong Nash equilibrium interchangeably. The same applies for the respective implementation concepts.

[4] In contrast with the case of no partial honesty, where NVP along with *Maskin-monotonicity* are sufficient. The well-known result is due to Maskin (1999).

[5] Jackson (1992) criticizes the use of canonical mechanisms in implementation theory as too permissive due to their unbounded strategy spaces. Instead, he derives a necessary condition for implementation with bounded mechanisms in undominated strategies. In the same context, Mukherjee et al. (2017) provide a full characterization when all agents are partially honest.

[6] Fault-tolerant Nash equilibrium was first introduced in the implementation literature by Eliaz (2002) as an equilibrium concept which is robust to the bounded rationality of a number of agents.

[7] Specifically, they show that the Walrasian correspondence, although it violates *Maskin-monotonicity* can be implemented via a market-type mechanism, where agents announce prices and consumption bundles. Like in the case of Kartik et al. (2014), no tail-chasing construction is used.

## 3. Preliminaries

Our society consists of a finite set of individuals $N = \{1, \ldots, n\}$ with $|N| = n \geq 3$. By $C \subseteq N$ we will denote a coalition of agents. The set of all possible social outcomes is denoted by $A$ and we typically assume that $|A| \geq 2$. Each agent $i$ is endowed with a preference ordering (complete, reflexive and transitive binary relation) over $A$ that is denoted by $R_i$. We denote the set of all such possible orderings for $i$ by $\mathcal{R}_i$ and, as usual, by $P_i$ and $I_i$ we denote the asymmetric and symmetric part of $R_i$, respectively. Define $\mathcal{R} \equiv \times_{i\in N}\mathcal{R}_i$ with a typical element $R = (R_1, \ldots, R_n)$ which we call a *preference profile* or simply, *state*. For each $i \in N$ let $L_i(a, R) = \{b \in A | aR_ib\}$ be agent $i$'s *lower contour set* of outcome $a$ in state $R$. A *social choice rule* (SCR) $f$ is a correspondence $f : \mathcal{R} \rightrightarrows A$ such that for all $R \in \mathcal{R}$, $\emptyset \neq f(R) \subseteq A$. A *social choice function* (SCF) is a single-valued SCR. For any $R \in \mathcal{R}$, we call $f(R)$ the set of $f$-optimal outcomes in state $R$.

A mechanism $G$ is a pair $(S, g)$, which consists of a strategy space $S = \times_{i\in N}S_i$, with $S_i$ being the set of available strategies for each $i \in N$, and an outcome function $g : S \to A$, that maps each strategy profile $s = (s_1, \ldots, s_n) \in S$ to an outcome in $A$. As usual, let $(s'_i, s_{-i})$ be the strategy profile where agent $i$ plays the strategy $s'_i$ while all $j \neq i$ play $s_j$. In a similar manner, let $(s'_C, s_{N\setminus\{C\}})$ be the strategy profile where all $i \in C$ play $s'_i$, and all $j \in N \setminus C$ play $s_j$. We also define the range of a mechanism $G$ as $g(S) = \{a \in A | a = g(s)$ for some $s \in S\}$. Now let $\Gamma$ be the set of all possible mechanisms, and $\Gamma^* = \{G \in \Gamma | g(S) = A\}$, that is, $\Gamma^*$ is the set of all mechanisms whose range is equal to the set of social outcomes. Any mechanism $G$ with a preference profile $R$ defines a normal form game $(G, R)$. We focus on the case of complete information where the state $R$ is common knowledge among the agents, while not to the planner.

In our setting, we assume that agents do not only care about the social outcomes, but also give some importance (although small) to the procedure that leads to those outcomes. More specifically, we assume that agents are *partially honest* in the following sense: If an agent is indifferent between two outcomes and she can attain those outcomes with two different strategies with one being "honest" and the other being "dis-honest", then she strongly prefers to follow the honest strategy. More formally, for honesty to be meaningful in our setting, we should restrict the set of possible mechanisms such that the strategy set of each $i \in N$ is $S_i = \mathcal{R} \times M_i$. That is, each agent is required to announce a preference profile $R \in \mathcal{R}$ and an arbitrary message $m_i \in M_i$. Then, given a mechanism $G$, for any $i \in N$ we define $i$'s *truthful correspondence* as $T_i^G : \mathcal{R} \rightrightarrows S_i$ such that for each agent $i$, state $R$ and message $m_i$, $T_i^G(R) = \{R\} \times M_i$. The truthful correspondence represents the truthful strategies for each agent $i$ in state $R$, which essentially consists of announcing the "true" state. We now define agent $i$'s *extended preferences* on the strategy space $S$ as follows. Given a vector of truthful correspondences $T^G = (T_1^G, \ldots, T_n^G)$, for all $i \in N$ and $R \in \mathcal{R}$, define $\succeq_i^R$ as a complete, transitive and reflexive binary relation on $S$. An extended preference profile in state $R$ is denoted by $\succeq^R = (\succeq_1^R, \ldots, \succeq_n^R)$. We are now ready to proceed to the formal definition of partial honesty.

Given a mechanism $G$, an agent $i$ is *partially honest* if $\forall s_i, s'_i \in S_i, \forall s_{-i} \in S_{-i}$:

- $[s_i \in T_i^G(R), s'_i \notin T_i^G(R)$ and $g(s_i, s_{-i})R_ig(s'_i, s_{-i})] \Rightarrow (s_i, s_{-i}) \succ_i^R (s'_{-i}, s_{-i})$.
- In all other cases, $g(s_i, s_{-i})R_ig(s'_i, s_{-i}) \iff (s_i, s_{-i}) \succeq_i^R (s'_{-i}, s_{-i})$

An agent $i$ is *not partially honest* if $\forall s_i, s'_i \in S_i, \forall s_{-i} \in S_{-i}$:

- $g(s_i, s_{-i})R_ig(s'_i, s_{-i}) \iff (s_i, s_{-i}) \succeq_i^R (s'_{-i}, s_{-i})$

In other words, an agent cares about honesty in a lexicographic manner: First she "consults" her ordering over outcomes, and if she is indifferent between some, she consults her ordering over strategies, strongly preferring the honest strategies if they exist. That is, her partial honesty serves the purpose of a *tie-breaking rule* when she faces indifference. On the other hand, an agent that is not partially honest cares only about the outcomes and does not give any significance to her strategies.

Notice that a mechanism $G$ with an extended preference profile $\succeq^R$ in state $R$ defines an (extended) game in normal form $(G, \succeq^R)$. Finally, we assume that in our society there can be partially honest and not partially honest agents and we denote the set of partially honest agents by $H$. For the planner however, we only assume that he knows the class of all conceivable sets of partially honest agents, $\mathcal{H} \subseteq 2^N \setminus \{\emptyset\}$, without knowing which set is the actual one.

Regarding the solution concept, since we assume that players are allowed to collude, the equilibrium notion that we use is *strong equilibrium*. Formally, $s \in S$ is a strong equilibrium in the game $(S, g, \succeq^R)$, if for all $C \subseteq N$ and $s'_C \in S_C$, there exists an agent $i \in C$ such that $(s_C, s_{N\setminus C}) \succeq_i^R (s'_C, s_{N\setminus C})$. In other words, a strategy profile is a strong equilibrium if there is no coalition that can deviate from it and make all of its members strictly better off. Let the set of strong equilibria of $(S, g, \succeq^R)$ be $SE(G, \succeq^R) = \{s \in S | s$ is a strong equilibrium in $(G, \succeq^R)\}$. We say that mechanism $G$ implements the SCR $f$ in strong equilibrium, if in any state $R \in \mathcal{R}$, $g(SE(G, \succeq^R)) = f(R)$, that is, if in any state, the set of outcomes obtained through the strong equilibria of the extended game coincides with the set of socially optimal outcomes. The SCR $f$ is strongly implementable if there exists a mechanism that implements it in strong equilibrium.

The previous formal setting can be interpreted as follows. First of all, the SCR represents the collective choice rule that our society utilizes in order to make collective decisions. It can also be interpreted as the constitution of the society designed in an *ex ante* stage. A mechanism on the other hand represents the institution through which the agents in the society interact with each other, that is, it determines the rules and the outcomes of the interaction. A hypothetical benevolent social planner wishes to implement the SCR, however, he does not know the true state, hence, he relies on the agents in order to obtain this information. On the other hand, truthful revelation of the state may not be in the best interests of some agents. Therefore, the goal of the social planner is to construct a mechanism that will lead to the optimal according to the SCR outcome, for any realization of the agents' preferences, that is, for any preference profile. For the strong implementation of the SCR we thus require any optimal outcome to be attainable by some strong equilibrium and any strong equilibrium to lead to an optimal outcome.

## 4. Results

In this section, we present our main results. Before proceeding though, it would be helpful first to review the result of Korpela (2013). This will enable us to outline more clearly the weakening of the sufficient conditions for strong implementation when we adopt the partial honesty assumption. The conditions are the following:

**Holocaust Alternative (HA)**: $\exists a_H \in A$, such that:

- $\forall R \in \mathcal{R}$, $a_H \notin f(R)$, and,
- $\forall R \in \mathcal{R}$, $\forall a \in A \setminus \{a_H\}$, $a \notin L_i(a_H, R)$.

**Weak Pareto Optimality (WPO)**: $\forall R \in \mathcal{R}, f(R) \subseteq wPO(A, R)$, where $wPO(A, R) = \{a \in A | \nexists b \in A$ such that $\forall i \in N, bP_ia\}$.

**Axiom of Sufficient Reason (ASR)**: $\forall R, R' \in \mathcal{R}, \forall a \in f(R), \forall b \in A$:

$\forall i \in N, L_i(a, R) \subseteq L_i(b, R') \Rightarrow b \in f(R')$.

Intuitively, one can imagine **HA** as the worst alternative for all agents in any state, that cannot ever be selected as an optimal outcome. It is a significant restriction on the preference domain, however, it is meaningful in various applications. It essentially allows us to overcome more involved conditions such as Condition $\gamma$ of Dutta and Sen (1991). **WPO** restricts the range of the SCR to weakly Pareto optimal outcomes. It is well-known from Maskin (1979) that weak Pareto optimality in the range of the mechanism is also a necessary condition for strong implementation.

**ASR** can be interpreted as follows: Let an outcome $a$ be selected as $f$-optimal for some preference profile $R$. Now imagine an outcome $b$ and profile $R'$ such that for all agents, every outcome that was ranked weakly below $a$ in $R$ is also ranked weakly below $b$ in $R'$. Then, $b$ should be $f$-optimal in $R'$. In other words, if every reason for $a$ to be $f$-optimal in $R$ is also a reason for $b$ to be $f$-optimal in $R'$, and $a$ is indeed selected as an optimal outcome in $R$, then $b$ should be selected as an optimal outcome in $R'$ as well. It is useful to note that **ASR** is stronger than *Maskin-monotonicity* (**MON**) and *Unanimity* (**U**) as it implies both. We review the formal definitions below:

**Maskin-Monotonicity (MON)**: $\forall R, R' \in \mathcal{R}, \forall i \in N, \forall a \in f(R)$:

$\forall i \in N, L_i(a, R) \subseteq L_i(a, R') \Rightarrow a \in f(R')$.

**Unanimity (U)**: $\forall R \in \mathcal{R}, \forall a \in A$:

$\forall i \in N, A \subseteq L_i(a, R) \Rightarrow a \in f(R)$.

For example, note that we obtain **MON** if in the definition of **ASR** we set $b = a$. To see that it implies **U**, suppose that **ASR** holds, and for some state $R$ and outcome $a$ we have that for all $i$, $A \subseteq L_i(a, R)$. Then, for any state $R'$ and any outcome $c \in f(R')$ it trivially holds that for all $i$, $L_i(c, R') \subseteq A \subseteq L_i(a, R)$, and from **ASR**, $a \in f(R)$ is obtained. We are now ready to present Korpela's theorem:

**Theorem 1** (*Korpela, 2013*)**.** *If a SCR $f$ satisfies **HA**, **WPO** and **ASR** then it is strongly implementable.*

Theorem 1 makes no assumptions with regards to the partial honesty motive. Its significance lies on the simplicity and intuitive appeal of the conditions. Now proceeding to our results, we will utilize the following assumption that summarizes the knowledge of the social planner with regards to the number of partially honest agents in the society.

**Assumption 1.** All agents in $N$ are partially honest and the planner knows that.

Assumption 1 has been extensively used in implementation problems. Examples include Kartik and Tercieux (2012), Korpela (2014), Matsushima (2008), Mukherjee et al. (2017), Ortner (2015) and Saporiti (2014). As in the case of the Dutta and Sen (2012) in Nash implementation, our goal is to examine the effect of the presence of partially honest agents on the strong implementation problem. Moreover, we aim to determine whether partial honesty bears analogous significant impact in the case of strong implementation as in Nash implementation, given that the sufficient conditions for the former are much stronger than in the case of the latter. In fact, by assuming that all agents are partially honest we manage to derive sharp and significant results. For our first result, we identify sufficient conditions for strong implementation when all agents are partially honest. Our key condition is the following[8]:

**Weak Pareto Dominance (WPD)**: $\forall R \in \mathcal{R}, \forall a \in f(R), \forall b \in A$, if:

- $\exists j \in N$, $aI_jb$, and
- $\forall i \in N \setminus \{j\}$, $bR_ia$,

then $b \in f(R)$.

The intuition behind our condition is the following: Suppose that $a$ is an $f$-optimal outcome at state $R$. Then, if there exists an outcome $b$ such that everyone weakly prefers $b$ to $a$, with at least one agent being indifferent between them, then $b$ must be selected as $f$-optimal as well.[9] Another way to look at **WPD** is as an "expansion" of the set of socially optimal outcomes in each state, so as to include all unanimously weakly preferred, or indifferent outcomes. The latter interpretation also has a strong normative appeal. Notice that **WPD** is implied by **ASR**. To see this simply set $R = R'$ in the definition of **ASR** which makes **WPD** true. Another interesting fact with regards to **WPD** is that together with **WPO**, it implies **U**, which will prove to be particularly useful in our main result. This is stated formally in Proposition 1.

**Proposition 1.** *If a SCR $f$ satisfies **WPO** and **WPD**, then it satisfies **U**.*

**Proof.** Consider a SCR $f$ that satisfies both **WPO** and **WPD**. Also, consider a state $R \in \mathcal{R}$ and an outcome $a \in A$ such that $\forall i \in N, \forall b \in A, aR_ib$, so that the premises of **U** are satisfied. If $a \in f(R)$, then we are done. Suppose that this is not the case. Then, since $f(R) \neq \emptyset$, there must exist an outcome $c \in A$ such that $c \in f(R)$. Since $\forall i \in N, \forall b \in A, aR_ib$, we must have that $aR_ic$. Now suppose that $\forall i \in N, aP_ic$. This however cannot be the case as **WPO** is violated. Therefore, there must exist an agent $j \in N$ such that $aI_jc$. However, for all $i \in N \setminus \{j\}$ it holds that $aR_ic$. But then, **WPD** implies $a \in f(R)$, a contradiction. This completes the proof. □

Next, we present the second part of our sufficient condition, a weakening of **HA**, the *Universally Worst Alternative*. It is particularly useful as it is satisfied in various interesting environments as shown in our applications section. We state it formally below:

**Universally Worst Alternative (UWA)**: $\exists a_W \in A$, such that $\forall R \in \mathcal{R}, \forall i \in N, \forall a \in f(R), aP_ia_W$.

So, a **UWA** is strictly worse than any socially optimal outcome for any agent and state and is never selected as socially optimal itself. It is easy to see that it is implied by **HA**, as any **HA** is also a **UWA**.[10] Now, **UWA**, **WPO** and **WPD** become sufficient for strong implementation when all agents are partially honest, which is stated in our main theorem:

**Theorem 2.** *Suppose that Assumption 1 holds. If a SCR $f$ satisfies **UWA**, **WPO**, and **WPD**, then it is strongly implementable.*

**Proof.** See Appendix. □

Regarding the proof, we utilize the mechanism of Korpela (2013). Each agent is called to announce an outcome, a state, a positive integer and whether she raises a flag or not. We essentially show that because of Assumption 1, there cannot be any strong equilibria where an agent is announcing a state different from the true one, as in such a case, due to the nature of the outcome function, there would exist profitable deviations motivated by partial honesty. Then, we show that our conditions are sufficient to guarantee that a socially optimal outcome is a strong equilibrium and that any strong equilibrium leads to a socially optimal outcome.

[8] We are grateful to an anonymous referee for motivating us to pursue a weakening of the condition that we initially presented in our working paper.

[9] In general, we can exclude the possibility of $a$ being strictly Pareto dominated by $b$ by the **WPO** condition which, apart from using it as part of our sufficient condition, we also show it to be necessary for partially honest strong implementation in the range of the mechanism. See Proposition 2.

[10] For other uses of **UWA** see Moore and Repullo (1990), or Jackson et al. (1994).

Several points are worth noting in this particular theorem. First, **WPD** constitutes a significant weakening of the **ASR** which reduces to a Pareto related condition. This is quite interesting since we were able to dispose of **MON**, or any variation of it from our sufficient conditions. In fact, we only utilize "intra-state" conditions, that is, conditions that restrict the socially optimal set with regards to the same state, rather than "inter-state" ones. The second point to note is that **WPO** is also a necessary condition for partially honest strong implementation, given that the range of the mechanism coincides with the set of alternatives.[11] We formally prove the statement in Proposition 2. Finally, notice that if we only allow for linear orderings,[12] **WPD** holds trivially (Proposition 3) and it becomes redundant as a sufficient condition. Below we provide the formal statements and appropriate proofs and in Corollary 1 we state a characterization theorem of strongly implementable SCRs for the case of linear preferences when agents are partially honest.

**Proposition 2.** *Let Assumption 1 hold and $f$ be strongly implementable by a mechanism $G \in \Gamma^*$. Then $f$ satisfies **WPO**.*

**Proof.** Let the premises hold. To derive a contradiction, suppose that $f$ does not satisfy **WPO**. This implies that for some $R \in \mathcal{R}$, there exists $a \in f(R)$ such that $a \notin wPO(A, R)$. So, there must exist $b \in A$ such that $\forall i \in N, bP_i a$. Now, since $f$ is strongly implementable, there exists a strong equilibrium $s \in S$ such that $g(s) = a$. So, $\forall C \subseteq N, \forall s'_C \in S_C, \exists j \in C, (s_C, s_{N\setminus C}) \succeq_j^R (s'_C, s_{N\setminus C})$. Since $G \in \Gamma^*$, we are allowed to consider $C = N$ and $g(s') = b$. Then, we have that $s \succeq_j^R s'$ and for $j$ it holds that:

- $s \sim_j^R s'$ (1), or
- $s \succ_j^R s'$ (2)

If (1) holds, then $g(s) = aI_j b = g(s')$, but also $bP_j a$, a contradiction. If (2) holds, we have either $g(s) = aP_j b = g(s')$ and $bP_j a$, a contradiction, or $a = g(s) I_j g(s')$, $s_i \in T_j^G(R)$ and $s'_i \notin T_j^G(R)$ which also contradicts $bP_j a$. So, our initial statement that $f$ does not satisfy **WPO** cannot hold. This completes the proof. □

**Proposition 3.** *If $\mathcal{R}^A = \mathcal{L}$, then any SCR $f$ satisfies **WPD**.*

The proof of Proposition 3 is straightforward, as one can notice that if there exists an agent that is indifferent between a socially optimal alternative $a$ and an outcome $b$, as dictated in the premise of **WPD**, then, by the linear preference assumption, $a$ must be equal to $b$ and the condition holds vacuously. We are now ready to proceed with our corollary:

**Corollary 1.** *Let $\mathcal{R}^A = \mathcal{L}$ and Assumption 1 hold. If a SCR $f$ satisfies **UWA**, then it is strongly implementable by a mechanism $G \in \Gamma^*$ if and only if it satisfies **WPO**.*

**Proof.** Immediate implication of Theorem 2 and Propositions 2 and 3. □

Corollary 1 provides a characterization of the strongly implementable social choice rules with linear preferences, when there exists a **UWA** and all agents are partially honest. Essentially, in this case **WPO** is a necessary and sufficient condition for strong implementation.[13]

[11] This assumption is crucial for the necessity of **WPO**.

[12] Formally, let $\mathcal{L}_i$ be the set of all linear, that is, complete, transitive and antisymmetric, orders on $A$ for each agent $i$ and let $\mathcal{L} \equiv \times_{i\in N} \mathcal{L}_i$. Let the space of admissible preferences be $\mathcal{R}^A$. So, in this case we set $\mathcal{R}^A = \mathcal{L}$.

[13] We thank an anonymous referee for pointing us to the possibility of this characterization theorem.

## 5. Applications

In this section we provide applications of our Theorem 2. Our first application is in pure matching environments, that is, one-to-one matching environments where for every agent, staying unmatched is not feasible, or it is the worst possible alternative in any state. For example, a manager in a firm might want to match people from two groups with different abilities in pairs, in order to undertake projects. In this case it might be reasonable to assume that staying unmatched is not feasible (as it might lead to redundancies). We show that when all agents are partially honest, the man-optimal (or woman-optimal) stable solution is strongly implementable. This is to be compared with the results of Tadenuma and Toda (1998), who show that with more than three agents in each group, while the whole stable solution in pure matching problems is Nash implementable, no single-valued subsolution of it is. Lombardi and Yoshihara (2017b) show that partial honesty can resolve this issue for Nash implementation, as the man-optimal (or woman-optimal) solution become Nash implementable in this case. With regards to strong implementation, Shin and Suh (1996) present a mechanism for strong implementation of the stable rule in one-to-one matching problems and the implementability of the stable rule in pure marriage problems is shown in Korpela (2013).

Our second application is in bargaining environments. We show that when all agents are partially honest, the Nash bargaining solution is strongly implementable. In general, it is known that the Nash bargaining solution is not Nash implementable, due to the result by Vartiainen (2007). However, Lombardi and Yoshihara (2017b) again show that it can be implemented with partial honesty. Our results extend theirs to the strong equilibrium concept.

### 5.1. Pure matching environments

We start by defining the formal pure matching environment. Let $M, W$ be two fixed finite sets, such that $|M| = |W| \geq 2$ and $M \cap W = \emptyset$. For all $i \in M$, $P_i$ is a linear order on $W \cup \{i\}$, and for all $i \in W$, $P_i$ is a linear order on $M \cup \{i\}$. A matching is a function $\mu : M \cup W \to M \cup W$ such that for any $i \in M \cup W$ the following hold:

- $i \in M \,\&\, \mu(i) \neq i \Rightarrow \mu(i) \in W$,
- $i \in W \,\&\, \mu(i) \neq i \Rightarrow \mu(i) \in M$, and
- $\mu(\mu(i)) = i$.

Let $\mathcal{M}$ be the set of all matchings. We now extend the relation $P_i$ to $\mathcal{M}$ by defining a new relation $R_i$ as follows:

$$\forall i \in M \cup W, \forall \mu, \mu' \in \mathcal{M}, \mu R_i \mu' \iff \mu(i) P_i \mu'(i) \text{ or } \mu(i) = \mu'(i)$$

Let the set of all preferences over $\mathcal{M}$ of each agent $i$ be $\mathcal{R}_i$. We then define $\mathcal{R} \equiv \times_{i\in M\cup W} \mathcal{R}_i$. As usual, $R \in \mathcal{R}$ denotes a preference profile. Now we make the following assumption, which makes our environment one of pure matching:

**Assumption 2.** $\forall m \in M, \forall w \in W, \forall \mu \in \mathcal{M}, wP_m m \,\&\, mP_w w$.

A solution (or SCR) is a correspondence $\varphi : \mathcal{R} \rightrightarrows \mathcal{M}$ such that for all $R \in \mathcal{R}$, $\varphi(R) \subseteq \mathcal{M}$. A pair $(m, w) \in M \times W$ *blocks* $\mu \in \mathcal{M}$ in $R \in \mathcal{R}$ if $wP_m\mu(m)$ and $mP_w\mu(w)$. A matching $\mu \in \mathcal{M}$ is *stable* in $R \in \mathcal{R}$, if there is no pair $(m, w) \in M \times W$ such that $(m, w)$ blocks $\mu$ in $R$. Let $S(R)$ be the set of all stable matchings in $R \in \mathcal{R}$. The stable matching rule is a rule $f^S : \mathcal{R} \rightrightarrows \mathcal{M}$ such that for every $R \in \mathcal{R}$, $f^S(R) = S(R)$. We say that $\mu^M \in \mathcal{M}$ is the *man-optimal* stable matching in state $R \in \mathcal{R}$ if $\mu^M \in S(R)$ and for every $\mu' \in S(R)$ and $m \in M$, we have that $\mu^M(m) P_m \mu'(m)$, or $\mu^M(m) = \mu'(m)$. The man-optimal stable rule $f^M$ is a function $f^M : \mathcal{R} \to \mathcal{M}$ such that for every $R \in \mathcal{R}$, $f(R) = \mu^M$. In a similar manner, we can define the woman-optimal stable matching and rule. We now proceed by stating our possibility result for the pure matching environment.

**Proposition 4.** *Let Assumptions 1 and 2 hold. Then, the man-optimal stable rule $f^M$ is strongly implementable.*

**Proof.** It suffices to show that $f^M$ satisfies **UWA**, **WPO** and **WPD**,

**Claim 1.** *$f^M$ satisfies **UWA**.*

**Proof.** By the construction of the pure matching environment, we have assumed that staying single is the worst alternative for every $i \in M \cup W$. So, we can set $a_W = \mu_W$, where for all $i \in M \cup W$, $\mu_W(i) = i$. So, our environment satisfies **UWA**.[14] □

**Claim 2.** *$f^M$ satisfies **WPO**.*

**Proof.** Suppose not. Consider $R \in \mathcal{R}$ such that $\mu = f^M(R)$ and suppose there exists $\mu' \in \mathcal{M}$ with $\mu' \neq \mu$ such that $\forall i \in M \cup W$, $\mu'(i) P_i \mu(i)$. Then, there exists $(m, w) \in M \times W$ such that $\mu'(m) = w \neq \mu(m)$ and $\mu'(w) = m \neq \mu(w)$. Consequently, the pair $(m, w)$ would block matching $\mu$, which contradicts its stability. Therefore, $f^M$ satisfies **WPO**. □

**Claim 3.** *$f^M$ satisfies **WPD**.*

**Proof.** Consider $R \in \mathcal{R}$ and let $f^M(R) = \mu^M$. Now suppose there exists $\mu \in \mathcal{M}$ such that:

- $\exists j \in N$, $\mu^M I_j \mu$, and
- $\forall i \in N \setminus \{j\}$, $\mu R_i \mu^M$

Since the man-optimal stable rule $f^M$ is a function, it suffices to show that $\mu = \mu^M$. Now, without loss of generality let $j = m \in M$. For $m$ it holds that $\mu I_m \mu^M$, which implies $\mu(m) = \mu^M(m)$. Let $\mu(m) = w$. Then necessarily it must be the case that $\mu(w) = \mu^M(w)$ and thus $\mu I_w \mu^M$. Now if for all $i \in M \cup W \setminus \{m, w\}$ it also holds that $\mu(i) = \mu^M(i)$, then $\mu = \mu^M$ and we are done. Suppose that this is not the case. So, there exists $i \in M \cup W \setminus \{m, w\}$ such that $\mu(i) \neq \mu^M(i)$. Again, without loss of generality, assume that $i = m' \in M$. Then, it must be that $\mu(m') P_{m'} \mu^M(m')$. Let $\mu(m') = w'$. Now, for $w'$ it is also true that $m' P_{w'} \mu^M(w')$. However, this contradicts the stability of the man-optimal stable matching $\mu^M$, as the couple $(m', w')$ would block it. Therefore, we conclude that $\mu = \mu^M$ and **WPD** holds. □

By Claims 1, 2, 3 and Theorem 2, we have that the man-optimal stable solution is strongly implementable. This completes the proof. □

### 5.2. Bargaining environments

For the definition of the bargaining environment we chose to follow the work of Vartiainen (2007), to whom we refer for the detailed formulation. Let $N = \{1, 2, \ldots, n\}$ be the set of players. The set of outcomes is $A = \{(a_1, \ldots, a_n) \in \mathbb{R}^n_+ | \sum_{i=1}^n a_i \leq 1\}$. Let the set of possible types of each agent $i \in N$ be $\Theta$. For each $\theta_i \in \Theta$, $v_i(\cdot, \theta_i) : [0, 1] \to \mathbb{R}$ is agent $i$'s strictly monotonic and continuous utility function. Let $\Theta_0$ be the normalized set of types for each $i$ such that $\Theta_0 = \{\theta_i \in \Theta | v_i(0, \theta_i) = 0\}$. Let $\Delta$ be the set of all probability distributions on $A$. So, for any outcome $p \in \Delta$ and agent $i \in N$, $v_i(p, \theta) = \int_A v_i(a_i, \theta_i) dp(a)$ is the utility function of $i$ defined on $\Delta$. We also set the disagreement points $\mathbf{d} = \mathbf{0}$. The Nash solution is a SCR $f^N : \Theta_0^n \rightrightarrows \Delta$ such that $\forall \theta \in \Theta_0^n$, $f^N(\theta) = \text{argmax}_{p \in \Delta} \Pi_{i=1}^n v_i(p, \theta_i)$. Notice that our environment satisfies **UWA**, since we have assumed strictly monotonic utility functions and in any Nash solution all agents get positive amounts of the good. This allows us to set $a_W = \mathbf{d} = \mathbf{0}$.

**Proposition 5.** *Let Assumption 1 hold. Then, the Nash solution $f^N$ is strongly implementable.*

**Proof.** Since the Nash solution satisfies weak Pareto optimality by definition, and our environment satisfies **UWA**, it suffices to show only that $f^N$ satisfies **WPD**.

**Claim 4.** *$f^N$ satisfies **WPD**.*

**Proof.** Consider $\theta \in \Theta_0^n$ such that $p \in f^N(\theta)$. Now, let $q \in \Delta$ be such that $\exists j \in N$, $v_j(q, \theta_j) = v_j(p, \theta_j)$ and $\forall i \in N \setminus \{j\}$, $v_i(q, \theta_i) \geq v_i(p, \theta_i)$. If $q = p$, then we are done. Suppose that $q \neq p$. If now for all $i \in N \setminus \{j\}$ it is also the case that $v_i(q, \theta_i) = v_i(p, \theta_i)$, then it must be that $q \in \text{argmax}_{p \in \Delta} \Pi_{i=1}^n v_i(p, \theta_i)$. Assume then that there exists an $i \in N \setminus \{j\}$ such that $v_i(q, \theta_i) > v_i(p, \theta_i)$. But this contradicts that $p \in \text{argmax}_{p \in \Delta} \Pi_{i=1}^n v_i(p, \theta_i)$. So, it is true that $f^N$ satisfies **WPD**. □

By Claim 4, Theorem 2 and the fact that the Nash bargaining solution satisfies **UWA** and **WPO**, we conclude that it is strongly implementable. This completes the proof. □

We have shown that the Nash solution satisfies our sufficient conditions and is thus strongly implementable when all agents are partially honest. For this result we relied on the ordinality of the environment. Note for example that **U** is not satisfied by the egalitarian solution in an environment where interpersonal comparisons are allowed, preferences are not strictly monotone and there is more than one good.[15] This implies that our Theorem 2 cannot be applied in this case.

## 6. Concluding remarks

We have provided a sufficiency theorem for strong implementation when all agents are partially honest. Our goal was to extend the positive results that have been obtained in partially honest Nash implementation to the solution concept of strong equilibrium. Our sufficient conditions are much stronger than in the case of Nash implementation and this is due to the much more demanding solution concept, as well as due to the attempt to provide simple sufficient conditions rather than a complete characterization.

As applications of our main theorem, we showed that the man-optimal (or woman-optimal) stable rule in a pure matching environment as well as the Nash solution in a bargaining environment with strictly monotone preferences are both strongly implementable when all agents are partially honest. However, as noted before, both these rules are not strongly implementable when there are no partially honest agents, therefore our results show the expansion of strongly implementable rules when the motive of minimal honesty is assumed.

In our view, the applications of our theorems provide an insight into the possibilities that arise in implementation theory when non-consequentialist motives are taken into account. They also emphasize the importance of procedural concerns in mechanism design and social choice theory. An interesting problem for further research which we aim to tackle, is closing the gap between our necessary and sufficient conditions. In fact, the *Non-emptiness* condition of Dutta and Sen (1991) is necessary in our case as well and we conjecture that it could constitute part of a sufficient condition, given that the mechanism is appropriately modified. In that way, the domain restriction of UWA could be avoided and more clear-cut results could be obtained. Finally, along the same line, it would be intriguing to study under which conditions partially honest strong implementation is equivalent to strong implementation.

[14] The pure matching environment actually satisfies the stronger condition **HA** as shown in Korpela (2013).

[15] For studies in bargaining theory in this type of environment see Roemer (1988).

**Acknowledgements**

The current research was conducted as part of my Ph.D. thesis at the University of Glasgow. I want to thank my supervisors, Michele Lombardi, Takashi Hayashi and Anna Bogomolnaia for their invaluable comments and support. I am also grateful to two anonymous referees whose comments greatly improved the results of the paper. Finally, I would like to thank Arunava Sen and the participants of the 14th meeting of the Society for Social Choice and Welfare for their comments and suggestions. This work was supported by the Economic and Social Research Council [ES/J500136/1].

## Appendix

*Mechanism*

For the proof of Theorem 2 we will utilize the following mechanism $G = (S, g)$:

For all $i \in N, S_i = A \times \mathcal{R} \times \{NF, F\} \times \mathbb{N}_+$. The outcome function $g$ is defined as follows:

(1) If $\forall i \in N, s_i = (a, R, NF, \cdot)$ and $a \in f(R)$, then $g(s) = a$.

(2) If $\exists C \subset N, \forall i \in N \setminus C, s_i = (a, R, NF, \cdot)$ with $a \in f(R)$, and $\forall j \in C, s_j = (a^j, R^j, F, n^j)$, then:

- If $k = \min\{\text{argmax}_{j\in C} n^j\}$ and $a^k \in \cup_{j\in C} L_j(a, R)$, then $g(s) = a^k$
- Otherwise, $g(s) = a$

(3) If $\forall i \in N, s_i = (a^i, R^i, F, n^i)$, then $k = \min\{\text{argmax}_{j\in N} n^j\}$ and set $g(s) = a^k$. (4) If none of the above apply, set $g(s) = a_W$.

**Proof of Theorem 2.** We will show that any SCR $f$ that satisfies our premises, namely **UWA**, **WPO** and **WPD** can be implemented by mechanism $G$ and we break the proof into two parts:

**Part 1:** $\forall R \in \mathcal{R}, f(R) \subseteq SE(R)$

Let the true state be $R^*$. Consider the strategy profile where $\forall i \in N, s_i = (a, R^*, NF, \cdot)$ and $a \in f(R^*)$. If $j \in N$ deviates to rule 2 she will obtain any $b \in L_j(a, R^*)$. So, $g(S_j, s_{N\setminus\{j\}}) = L_j(a, R^*)$. If any $C \subset N$ deviates to rule 2, the obtained outcome will be in $L_j(a, R^*)$ for at least one $j \in C$. If $N$ deviate to rule 3, there cannot be an improvement for all $i \in N$ since $f$ satisfies **WPO**. Finally, there is no profitable deviation by any coalition to rule 4, since, by definition of the **UWA**, $a_W$ is ranked strictly worse to any socially optimal outcome, by all agents. Therefore, $s$ is a strong equilibrium in $R^*$.

**Part 2:** $\forall R \in \mathcal{R}, SE(R) \subseteq f(R)$.

Let the true state be $R^*$. We proceed by first proving three useful claims:

**Claim 1*.** *There is no strong equilibrium under rule 1 where* $\forall i \in N, R^i \neq R^*$.

**Proof.** Suppose there exists a strong equilibrium under rule 1, where $\forall i \in N, s_i = (a, R, NF, \cdot)$ with $a \in f(R)$ and $R \neq R^*$. By rule 1 the outcome is $a$. Then, $\forall i \in N, s_i \notin T_i^G(R^*)$, so, any $i \in N$ can deviate to $s_i' = (a, R^*, F, n^i) \in T_i^G(R^*)$ inducing rule 2 while announcing the true state and not changing the outcome. Therefore, $s$ cannot be a strong equilibrium. □

**Claim 2*.** *There is no strong equilibrium under rule 2 where* $\exists i \in N \setminus C$ *such that* $R^i \neq R^*$.

**Proof.** Suppose there exists a strong equilibrium under rule 2 where $\exists i \in N \setminus C, s_i = (a, R, NF, \cdot)$ with $a \in f(R), R \neq R^*$, and $\forall j \in C, s_j = (a^j, R^j, F, n^j)$ and let $g(s) = b$. Then, we have that $s_i \notin T_i^G(R^*)$. We break the proof into two cases:

**Case 1**: $|N \setminus C| \geq 2$

- If $b = a$: Then, since by definition $a \in L_i(a, R)$ holds, $i$ can play $s_i' = (a, R^*, F, n^i) \in T_i^G(R^*)$ with a sufficiently high integer without changing the outcome and become strictly better off by Rule 2.
- If $b \neq a$: Then, again, since $b \in \cup_{j\in C} L_j(a, R)$ it must hold that $b \in \cup_{j\in C\cup\{i\}} L_i(a, R)$, so agent $i$ can play $s_i' = (b, R^*, F, n^i) \in T_i^G(R^*)$ with a sufficiently high integer without changing the outcome and become strictly better off by Rule 2.

**Case 2**: $N \setminus C = \{i\}$

In this case $i$ can play $s_i' = (b, R^*, F, n^i) \in T_i^G(R^*)$ with a sufficiently high integer without changing the outcome and become strictly better off by Rule 3.

Therefore, there is no strong equilibrium under rule 2, where for some $i \in N \setminus C, R^i \neq R^*$. □

**Claim 3*.** *There is no strong equilibrium under rule 2 where* $\exists i \in C$, *with* $R^i \neq R^*$.

**Proof.** Suppose this is not the case, that is, there exists a strong equilibrium under rule 2 such that $\exists i \in C$, with $R^i \neq R^*$. Also, by Claim 2*, we have established that in any strong equilibrium that falls in Rule 2, $\forall j \in N \setminus C, R^j = R^*$. So, we consider a case where $\forall j \in N \setminus C, s_j = (a, R^*, NF, \cdot)$ with $a \in f(R^*)$ and $\forall k \in C, s_k = (a^k, R^k, F, n^k)$ such that $R^k \neq R^*$ for some $i \in C$, that is, $\exists i \in C$ such that $s_i \notin T_i^G(R^*)$. Moreover, let $g(s) = b$. Now we take two mutually exclusive cases:

**Case 1**: $|C| \geq 2$

- If $b = a$, then, since we have that $a \in L_i(a, R^*)$ by definition, agent $i$ can play $s_i' = (a, R^*, F, n^i) \in T_i^G(R^*)$ with a sufficiently high $n^i$ inducing rule 2 without changing the outcome and becoming strictly better off.
- If $b = a^l \neq a$, where $l = \min\{\text{argmax}_{j\in C} n^j\}$, we distinguish two cases:
  - $l \neq i$: In this case, since $a^l \in \cup_{j\in C} L_j(a, R^*)$, agent $i$ can deviate to $s_i' = (b, R^*, F, n^i) \in T_i^G(R^*)$, win the integer game for a sufficiently high integer without affecting the outcome, and thus become better off by rule 2.
  - $l = i$: Again, $a^l \in \cup_{j\in C} L_j(a, R^*)$, so $i$ can play $s_i' = (b, R^*, F, n^i) \in T_i^G(R^*)$ and again become better off by rule 2.

**Case 2**: $C = \{i\}$.

- If $b = a$, then $i$ can deviate to $s_i' = (a, R^*, NF, \cdot) \in T_i^G(R^*)$ inducing Rule 1 and become better off by announcing the truth.
- If $b \neq a$, then it must be that $b = a^i$. So, since $b \in L_i(a, R^*)$, $i$ can revert to truth-telling by playing $s_i' = (b, R^*, F, n^i) \in T_i^G(R^*)$ and become better off by rule 2.

Therefore, there is no strong equilibrium under rule 2 where $\exists i \in C$ such that $R^i \neq R^*$. □

**Claim 4*.** *There is no strong equilibrium under rule 3 where* $\exists i \in N$, *with* $R^i \neq R^*$.

**Proof.** Suppose there exists a strong equilibrium under rule 3 where $\forall j \in N, s_j = (a^j, R^j, F, n^j), g(s) = b$ and let $R^i \neq R^*$ for some $i \in N$, that is, $\exists i \in N$ such that $s_i \notin T_i^G(R^*)$. Then, $i$ can deviate to $s_i' = (b, R^*, F, n^i) \in T_i^G(R^*)$ and obtain $b$ while announcing the true state $R^*$, for a sufficiently high integer $n^i$. Therefore, $s$ cannot be a strong equilibrium. □

**Claim 5*.** *There is no strong equilibrium under rule 4.*

**Proof.** Suppose on the contrary that there exists one, namely $s \in S$, with $g(s) = a_W$. So, $\forall C \subseteq N, \forall s'_C \in S_C, \exists i \in C, (s_C, s_{N\setminus C}) \succeq_i^{R^*} (s'_C, s_{N\setminus C})$. Consider the case where $C = N$ and let $g(s') = a \in f(R^*)$. Then, there exists $i \in N$ such that:

- $(s_C, s_{N\setminus C}) \succ_i^R (s'_C, s_{N\setminus C})$ (1), or
- $(s_C, s_{N\setminus C}) \sim_i^R (s'_C, s_{N\setminus C})$ (2).

Suppose (1) holds. Then, either $g(s) = a_W P_i a = g(s') \in f(R^*)$, which is a contradiction of **UWA**, or $g(s) = a_W I_i^* a = g(s') \in f(R^*)$, $s_i \in T_i^G(R^*)$ and $s'_i \notin T_i^G(R^*)$, where we have a contradiction as well. If (2) holds, then $g(s) = a_W I_i^* a = g(s') \in f(R^*)$ and the same contradiction emerges. So, there is no strong equilibrium under rule 4 and this completes the proof. □

**Corollary 2.** *Any strong equilibrium s of the mechanism G, falls under rules 1–3 and it also holds that* $\forall i \in N, R^i = R^*$.

**Proof.** Immediate implication of Claims 1*–5*. □

By the above arguments, we can restrict attention to strong equilibria under rules 1, 2 or 3, where $\forall i \in N, R^i = R^*$. Consider a strong equilibrium under rule:

1. That is, $\forall i \in N, s_i = (a, R^*, NF, \cdot)$. Then $g(s) = a \in f(R^*)$.
2. That is, $\forall i \in N \setminus C, s_i = (a, R^*, F, \cdot)$ with $a \in f(R^*)$, and $\forall j \in C, s_j = (a^j, R^*, F, n^j)$. Let $g(s) = b$. We distinguish two cases: $|N \setminus C| \geq 2$: Then, it must be that $\forall i \in N \setminus C, g(S_i, s_{N\setminus\{i\}}) = \cup_{j\in C\cup\{i\}} L_j(a, R^*)$ and $\forall j \in C, g(S_j, s_{N\setminus\{j\}}) = \cup_{j\in C} L_j(a, R^*)$, from Rule 2. For $s$ to be a strong equilibrium, it must hold that $\forall i \in N \setminus C, L_i(a, R^*) \subseteq \cup_{j\in C\cup\{i\}} L_j(a, R^*) \subseteq L_i(b, R^*)$ and, $\forall j \in C, L_j(a, R^*) \subseteq \cup_{j\in C} L_j(a, R^*) \subseteq L_j(b, R^*)$. So, for any $i \in N$ we have that $L_i(a, R^*) \subseteq L_i(b, R^*)$. However, since $a \in f(R^*)$, from **WPO**, it cannot be the case that $\forall i \in N, bP_i^* a$. So there must exist $j \in N$ such that $aI_j^* b$. From **WPD** it follows that $b \in f(R^*)$. $N \setminus C = \{i\}$: Then, for $i$ it must hold that $g(S_i, s_{N\setminus\{i\}}) = A$ from rule 3, and $\forall j \in C$ it must hold that $g(S_j, s_{N\setminus\{j\}}) = \cup_{j\in C} L_j(a, R^*)$ by rule 2. For $s$ to be a strong equilibrium, it must hold that $\forall i \in N \setminus C, L_i(a, R^*) \subseteq A \subseteq L_i(b, R^*)$ and $\forall j \in C, L_j(a, R^*) \subseteq \cup_{j\in C} L_j(a, R^*) \subseteq L_j(b, R^*)$. So for all $i \in N$ it holds that $L_i(a, R^*) \subseteq L_i(b, R^*)$. As before, from **WPO** and the fact that $a \in f(R^*)$, there must exist $j \in N$ such that $aI_j^* b$. Again, from **WPD** we must have that $b \in f(R^*)$.
3. That is, $s_i = (a^i, R^*, F, n^i), \forall i \in N$ and let $g(s) = b$. Then, $\forall i \in N$, it must hold that $g(S_i, s_{N\setminus\{i\}}) = A$. Now, for $s$ to be a strong equilibrium it must be that $\forall i \in N, A \subseteq L_i(b, R^*)$. Then, from **WPO**, **WPD** and Proposition 1, it must hold that $b \in f(R^*)$.

This completes the proof.

## References


Adachi, T., 2017. Nash Implementation with Honesty and Ranges of Honesty. mimeo.

Aumann, R.J., 1960. Acceptable points in games of perfect information. Pacific J. Math. 10 (2), 381–417.

Bierbrauer, F., Netzer, N., 2016. Mechanism design and intentions. J. Econom. Theory 163, 557–603.

Corchón, L.C., Herrero, C., 2004. A decent proposal. Spanish Econ. Rev. 6 (2), 107–125.

Doğan, B., 2017. Eliciting the socially optimal allocation from responsible agents. J. Math. Econom. 73, 103–110.

Doghmi, A., Ziad, A., 2013. On partially honest Nash implementation in private good economies with restricted domains: A sufficient condition. BE J. Theoret. Econ. 13 (1), 415–428.

Dutta, B., Sen, A., 1991. Implementation under strong equilibrium: A complete characterization. J. Math. Econom. 20 (1), 49–67.

Dutta, B., Sen, A., 2012. Nash implementation with partially honest individuals. Games Econom. Behav. 74 (1), 154–169.

Eliaz, K., 2002. Fault tolerant implementation. Rev. Econom. Stud. 69 (3), 589–610.

Glazer, J., Rubinstein, A., 1998. Motives and implementation: On the design of mechanisms to elicit opinions. J. Econom. Theory 79 (2), 157–173.

Hagiwara, M., 2017. Double Implementation with Partially Honest Agents. mimeo.

Hagiwara, M., Yamamura, H., Yamato, T., 2018. Implementation with socially responsible agents. Econ. Theory Bull. 55.

Hurkens, S., Kartik, N., 2009. Would I lie to you? On social preferences and lying aversion. Exp. Econ. 12 (2), 180–192.

Jackson, M.O., 1992. Implementation in undominated strategies: A look at bounded mechanisms. Rev. Econom. Stud. 59 (4), 757–775.

Jackson, M.O., 2001. A crash course in implementation theory. Soc. Choice Welf. 18 (4), 655–708.

Jackson, M.O., Palfrey, T.R., Srivastava, S., 1994. Undominated nash implementation in bounded mechanisms. Games Econom. Behav. 6 (3), 474–501.

Kartik, N., Tercieux, O., 2012. Implementation with evidence. Theoret. Econ. 7 (2), 323–355.

Kartik, N., Tercieux, O., Holden, R., 2014. Simple mechanisms and preferences for honesty. Games Econom. Behav. 83, 284–290.

Kimya, M., 2017. Nash implementation and tie-breaking rules. Games Econom. Behav. 102, 138–146.

Korpela, V., 2013. A simple sufficient condition for strong implementation. J. Econom. Theory 148 (5), 2183–2193.

Korpela, V., 2014. Bayesian implementation with partially honest individuals. Soc. Choice Welf. 43 (3), 647–658.

Lombardi, M., Yoshihara, N., 2017a. Natural implementation with semi-responsible agents in pure exchange economies. Internat. J. Game Theory 1–22.

Lombardi, M., Yoshihara, N., 2017b. Partially honest nash implementation: A full charcterization. Available at SSRN: https://ssrn.com/abstract=2232274 or http://dx.doi.org/10.2139/ssrn.2232274.

Lombardi, M., Yoshihara, N., 2018. Treading a fine line: (im)possibilities for Nash implementation with partially-honest individuals. Games Econom. Behav. 111, 203–216.

Maskin, E., 1979. Implementation and strong nash equilibrium. In: Laffont, J.J. (Ed.), Aggregation and Revelation of Preferences. North Holland, pp. 433–440.

Maskin, E., 1999. Nash equilibrium and welfare optimality. Rev. Econom. Stud. 66 (1), 23–38.

Matsushima, H., 2008. Role of honesty in full implementation. J. Econom. Theory 139 (1), 353–359.

Moore, J., Repullo, R., 1990. Nash implementation: a full characterization. Econometrica 1083–1099.

Moulin, H., Peleg, B., 1982. Cores of effectivity functions and implementation theory. J. Math. Econom. 10 (1), 115–145.

Mukherjee, S., Muto, N., Ramaekers, E., 2017. Implementation in undominated strategies with partially honest agents. Games Econom. Behav. 104, 613–631.

Ortner, J., 2015. Direct implementation with minimally honest individuals. Games Econom. Behav. 90, 1–16.

Roemer, J.E., 1988. Axiomatic bargaining theory on economic environments. J. Econom. Theory 45 (1), 1–31.

Saporiti, A., 2014. Securely implementable social choice rules with partially honest agents. J. Econom. Theory 154, 216–228.

Shin, S., Suh, S.-C., 1996. A mechanism implementing the stable rule in marriage problems. Econom. Lett. 51 (2), 185–189.

Suh, S.-C., 1996. Implementation with coalition formation: A complete characterization. J. Math. Econom. 26 (4), 409–428.

Suh, S.-C., 1997. Double implementation in nash and strong Nash equilibria. Soc. Choice Welf. 14 (3), 439–447.

Tadenuma, K., Toda, M., 1998. Implementable stable solutions to pure matching problems. Math. Social Sci. 35 (2), 121–132.

Vartiainen, H., 2007. Nash implementation and the bargaining problem. Soc. Choice Welf. 29 (2), 333–351.