Strong implementation with partially honest individuals

Foivos Savva

Adam Smith Business School, University of Glasgow

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 favor 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 a weaker version of the Holocaust Alternative, the Universally Worse Alternative. We finally provide applications of our result in pure matching and bargaining environments.

JEL classification: D71, D78.

Keywords: Strong Implementation, Partial Honesty, Tie-breaking rule

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

\(^1\)For a comprehensive survey of the main results in the literature of implementation theory see Jackson (2001).
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. In
the field of mechanism design this weak preference for honesty is usually modelled
in the following manner: Suppose that an agent is indifferent between two out-
comes. Then she will strictly prefer to send a truthful message rather than 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 semi-
al paper, Dutta and Sen (2012) show that in the presence of just one partially
honest agent in the society, Maskin monotonicity is no longer necessary condition
for Nash implementation and No Veto Power alone becomes sufficient for three or
more agents.

Overall, the results on Nash implementation with partial honesty have been
positive. An important question that remains 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\(^3\) a 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

\(^2\)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.

\(^3\)From now on we will use the terms strong equilibrium and strong Nash equilibrium inter-
changably. The same applies for the respective implementation concepts.
choice rule satisfies *Weak Pareto Optimality* (WPO), *Universally Worse 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 that if an outcome is optimal at some state, and if there is another outcome that is weakly preferable by all agents in the same state, then the latter should 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.

\(^4\)In contrast with the case of no partial honesty, where NVP along with *Maskin monotonicity* are sufficient. The famous result is due to Maskin (1999).
for both unanimous and non-unanimous social choice rules. 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\(^5\). 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, Lombardi and Yoshihara (2017\(^c\)) explore the possibility of strategy space reduction in partially honest Nash implementation and in Lombardi and Yoshihara (2017\(^c\)) they explore under which conditions partially honest Nash implementation is equivalent to Nash implementation.

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 (2017\(^a\)), the effect of social responsibility is explored with regards to natural implementation\(^7\). Hagiwara et al. (2017) utilize a similar concept of social responsibility for strategy space reduction with an outcome mechanism for Nash implementation. In a different environment, Dogan (2017) shows that the

\(^5\)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 and Muto (2016) 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.
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. 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 $R_i$ and, as usual, by $P_i$ and $I_i$ we denote the asymmetric and symmetric part of $R_i$ respectively. Define $R \equiv \times_{i \in N} R_i$ 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 $L_i(a, R) = \{b \in A | aR_i b\}$ be agent $i$'s lower contour set of outcome $a$ in state $R$. A Social Choice Rule (SCR) $f$ is a correspondence $f : R \Rightarrow A$ such that for all $R \in R$, $\emptyset \neq f(R) \subseteq A$. A Social Choice Function (SCF) is a single-valued SCR. For any $R \in R$, 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 $g : S \rightarrow A$, 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 $s_j$. Any mechanism $G$ with a preference profile $R$ define 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, in order for honesty to be meaningful in our setting, we define the strategy set of each $i \in N$ to be $S_i = R \times M_i$. That is, each agent is required to announce a preference profile $R \in R$ 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 $T^G_i : R \Rightarrow S_i$ such that for each agent $i$ and state $R$, $m_i$, $T^G_i(R) = \{R\} \times M_i$.
The truthful correspondence represents the truthful strategies for each agent \( i \) in state \( R \), which essentially consist 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 \( T^G = (T^G_1, \ldots, T^G_n) \), for all \( i \in N \) and \( R \in \mathcal{R} \), define \( \succeq^R_i \) as a complete, transitive and reflexive binary relation on \( S \). An extended preference profile in state \( R \) is denoted by \( \succeq^R = (\succeq^R_1, \ldots, \succeq^R_n) \). We are now ready to proceed to the formal definition of partial honesty.

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

- \([s_i \in T^G_i(R), s'_i \notin T^G_i(R) \text{ and } g(s_i, s_{-i}) R_i g(s'_i, s_{-i})] \Rightarrow (s_i, s_{-i}) \succeq^R_i (s'_i, s_{-i}).\)

- In all other cases, \( g(s_i, s_{-i}) R_i g(s'_i, s_{-i}) \iff (s_i, s_{-i}) \succeq^R_i (s'_i, s_{-i}).\)

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

- \( g(s_i, s_{-i}) R_i g(s'_i, s_{-i}) \iff (s_i, s_{-i}) \succeq^R_i (s'_i, s_{-i}).\)

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 significance to her strategies.

Notice that a mechanism \( G \) with an extended preference profile \( \succeq^R \) in state \( R \) define 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 \( g(s) \succeq^R_i g(s_C', s_{N\setminus C}) \). 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 \( SE(G, \succeq^R) = \{s \in S|s \text{ is a strong equilibrium in } (G, \succeq^R)\} \). We say that the mechanism \( G \) implements the SCR \( f \) in strong equilibrium, if in any state \( R \in \mathcal{R} \), \( SE(G, \succeq^R) = f(R) \). 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 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 $wPO(A, R) = \{a \in A| \exists b \in A \text{ such that } \forall i \in N, bP_i a\}$.

**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, **HA** can be thought of as a worse alternative for all agents, that cannot ever be selected as an optimal outcome. It is a significant restriction on the preference domain, however, it can be meaningful in some applications. It essentially allows us to overcome more involved assumptions regarding the intersection of the lower contour sets, as in the non-emptiness condition of Dutta and Sen (1991). **WPO** restricts the range of the SCR to weakly Pareto optimal outcomes. Note that weak Pareto optimality 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 \( b \) is weakly more preferable to \( a \) in \( R' \) by all agents. 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 \( \text{ASR} \) is stronger than \( \text{Maskin monotonicity (MON)} \) and \( \text{Unanimity (U)} \) 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 \( \text{MON} \) if in the definition of \( \text{ASR} \) we set \( b = a \). To see that it implies \( \text{U} \), suppose that \( \text{ASR} \) 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 \( \text{ASR} \), \( a \in f(R) \) obtains. We are now ready to present Korpela’s theorem:

**Theorem 1** (Korpela, 2013). If a SCR \( f \) satisfies \( \text{HA, WPO} \) and \( \text{ASR} \) 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 assumptions, which summarize the knowledge of the social planner regarding the number of partially honest agents in the society.

**Assumption 1:** All agents in \( N \) are partially honest and the planner knows that.

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 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. For our first result, we identify sufficient conditions for strong implementation when all agents are partially honest. Our key condition is the following:

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

\[ L_i(a, R) \subseteq L_i(c, R), \forall i \in N \Rightarrow c \in f(R) \]
The intuition behind our condition is the following: Suppose that \(a\) is an \(f\)-optimal outcome at state \(R\). Then, let \(b\) be an outcome such that, either it weakly Pareto dominates \(a\), or all agents are indifferent between \(a\) and \(b\), at the same state \(R\). Then, \(b\) must be selected as \(f\)-optimal as well. Notice that we can exclude the possibility of \(a\) being strictly Pareto dominated by \(b\) by the WPO condition. Finally, notice that WPD is implied by ASR and implies \(U^{8}\).

Next, we present a weakening of HA, the Universally Worse Alternative. It is particularly useful as it is satisfied in various interesting environments. We state it formally below:

Universally Worse Alternative (UWA): \(\exists a_W \in A\), such that \(\forall R \in \mathcal{R}, \forall i \in N, \forall a \in f(R), aP_i a_W\)

So, a UWA is strictly worse than any socially optimal outcome 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\(^9\). 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 WPO, HA, WPD, then it is strongly implementable.

*Proof.* See appendix.

Two comments are worth noting in this particular theorem. First, WPD constitutes a significant weakening of the ASR which reduces to a Pareto related condition. The second point to note is that if we only allow for linear orderings\(^{10}\), then, WPD is implied by WPO. Therefore, WPD becomes redundant as a sufficient condition. Below we provide a formal proof and in Corollary 1 we state the sufficiency theorem for the case of linear preferences.

**Proposition 1.** If \(\mathcal{R}^A = \mathcal{L}\), then WPO implies WPD.

*Proof.* Suppose that \(\mathcal{R}^A = \mathcal{L}\) and assume that WPO holds. Now consider \(R \in \mathcal{R}\), \(c \in A\) and \(a \in f(R)\) such that \(\forall i \in N, L_i(a, R) \subseteq L_i(c, R)\). We distinguish two cases:

- \(a = c\): Then, \(c \in f(R)\) and WPD holds.

---

\(^8\)To see that it is implied by ASR simply set \(R = R'\) in the definition of ASR. To show that it implies U, a similar argument as in the ASR\(\Rightarrow U\) implication applies.

\(^9\)For other uses of UWA see Jackson (2001), p. 686.

\(^{10}\)Formally, let \(L_i\) be the set of all linear, that is, complete, transitive and antisymmetric, orders on \(A\) for each agent \(i\) and let \(\mathcal{L} = \times_{i \in N} L_i\). Let the space of admissible preferences be \(\mathcal{R}^A\). So, in this case we set \(\mathcal{R}^A = \mathcal{L}\).
• $a \neq c$: Then, since we have linear preferences, $\forall i \in N, L_i(a, R) \subseteq L_i(c, R)$ implies that $\forall i \in N, cP_i a$. But, because $f$ satisfies WPO, this contradicts that $a \in f(R)$, so it cannot be the case.

\[\square\]

**Corollary 1.** Let $\mathcal{R}^A = \mathcal{L}$ and assumption 2 hold. If a SCR $f$ satisfies WPO and HA, then it is strongly implementable.

**Proof.** Immediate implication of Theorem 2 and Proposition 1. \[\square\]

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 a feasible alternative, or it is the worst alternative. For example, a manager in a firm might want to match people from two groups in pairs, 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) construct 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 implementation 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:
\[ \begin{align*}
\bullet & \quad i \in M \land \mu(i) \neq i \Rightarrow \mu(i) \in W, \\
\bullet & \quad i \in W \land \mu(i) \neq i \Rightarrow \mu(i) \in M, \text{ and} \\
\bullet & \quad \mu(\mu(i)) = i.
\end{align*} \]

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 \( R_i \). We then define \( R = \times_{i \in M \cup W} R_i \). As usual, \( R \in R \) denotes a preference profile. Now we make the following assumption, which makes our environment one of pure matching:

**Assumption 3**: \( \forall m \in M, \forall w \in W, \forall \mu \in \mathcal{M}, w P_m m \land m P_w w \).

A solution (or SCR) is a correspondence \( \varphi : R \Rightarrow \mathcal{M} \) such that for all \( R \in R, \varphi(R) \subseteq \mathcal{M} \). A pair \((m, w) \in M \times W\) blocks \( \mu \in \mathcal{M} \) in \( R \in R \) if \( w P_m \mu(m) \) and \( m P_w \mu(w) \). A matching \( \mu \in \mathcal{M} \) is stable in \( R \in R \), 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 \( R \in R \). The stable matching rule is a rule \( f^S : R \Rightarrow \mathcal{M} \) such that for every \( R \in R \), \( f^S(R) = S(R) \). We say that \( \mu^M \in \mathcal{M} \) is the man-optimal stable matching in state \( R \in R \) if \( \mu^M \in S(R) \) and for every \( \mu' \in S(R) \) and \( m \in M \), we have that \( \mu(m) P_m \mu'(m) \). The man-optimal stable rule \( f^M \) is a function \( f^M : R \rightarrow \mathcal{M} \) such that for every \( R \in R \), \( f(R) = \mu^M \). 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 2**. Let Assumptions 1 and 3 hold. Then, the man-optimal stable rule \( f^M \) is strongly implementable.

**Proof**. It suffices to show that \( f^M \) satisfies \textbf{WPO}, \textbf{WPD}, \textbf{UWA}.

**Claim 1**: \( f^M \) satisfies \textbf{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 \textbf{UWA} (the pure matching environment actually satisfies the stronger condition \textbf{HA} as show in Korpela (2013)).

**Claim 2**: \( f^M \) satisfies \textbf{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 the matching $\mu$ since we assumed that $\forall i \in M \cup W, \mu'(i)P_i \mu(i)$. This however contradicts the stability of $\mu$. Therefore, $f^M$ satisfies WPO.

\[\square\]

Claim 3: $f^M$ satisfies WPD.

Proof. Consider $R \in \mathcal{R}$ and let $f^M(R) = \mu^M$. Since the man-optimal stable rule $f^M$ is a function, it suffices to show that for any $\mu \in \mathcal{M}$, $[\forall i \in M \cup W, L_i(\mu^M, R) \subseteq L_i(\mu, R)] \Rightarrow \mu = \mu^M$. Suppose not. That is, suppose that there exists $\mu \neq \mu^M$ such that for all $i \in M \cup W, L_i(\mu^M, R) \subseteq L_i(\mu, R)$. Then, because $\mu \neq \mu^M$, there must exist $(m, w) \in M \cup W$, such that $\mu(m)P_m \mu^M(m)$ and $\mu(w)P_w \mu^M(w)$. This however contradicts the stability of $\mu^M$. So $f^M$ satisfies WPD.

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

\[\square\]

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 $N = \{1, 2, ..., n\}$ be the set of players. The set of outcomes is $A = \{(a_1, ..., a_n) \in \mathbb{R}^n_+ | \sum_{i=1}^n a_i \leq 1\}$. Let the set of possible types of each agent $i \in N$ be $\Theta_i$. For each $\theta_i \in \Theta_i$, $v_i(\cdot, \theta_i) : [0, 1] \rightarrow \mathbb{R}$ is agent $i$'s strictly monotonic and continuous utility function. Let $\Theta_0$ be the normalized set of types for each $i$ such that $\Theta_0 = \{\theta_i \in \Theta_i | v_i(0, \theta_i) = 0\}$. 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 $d = 0$. The Nash solution is a SCF $f^N : \Theta_0^N \rightarrow \Delta$ such that $\forall \theta \in \Theta_0^N$, $f^N(\theta) = \text{argmax}_{p \in \Delta} \prod_{i=1}^n v_i(p, \theta_i)$. Notice that our environment satisfies UWA, since we have assumed strictly monotonic utility functions, so we can take $a_H = 0$.

Proposition 3. 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: $f^N$ satisfies WPD.
Proof. Consider $\theta \in \Theta_0^n$ such that $p = f^N(\theta)$. Now, let $q \in \Delta$ be such that $L_i(p, \theta_i) \subseteq L_i(q, \theta_i), \forall i \in N$. Then, it must be that $\forall i \in N, v_i(q, \theta_i) \geq v_i(p, \theta_i)$. Since $f^N$ is a function, we need to show that $q = p$. Suppose not. That is, $\exists j \in N$ such that $v_j(q, \theta_j) > v_j(p, \theta_j)$. But this contradicts the fact that $f^N$ is Pareto optimal. Therefore, $p = q = f^N(\theta)$. So $f^N$ satisfies WPD. 

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. However, note that WPD is not satisfied by the egalitarian solution when preferences are not strictly monotone and there is more than one good. This is due to the fact that the egalitarian solution in this case does not satisfy $U$ which is implied by WPD.

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. Further research on the issue could potentially include a full characterization of partially honest strong implementation. In that way, the domain restriction of HA could be avoided and more clear-cut results could be obtained. Finally, it would be interesting to study under which conditions partially honest strong implementation is equivalent to strong implementation.
Acknowledgements

The current research was conducted as part of my PhD 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. This work was supported by the Economic and Social Research Council.

Appendix

Mechanism

For the proof of Theorem 2 we will utilize the following mechanism $G$:

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

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

2) If $\exists C \subset \mathbb{N}, \forall i \in \mathbb{N} \setminus C, s_i = (a, R, NF, \cdot)$ with $a \in f(R)$, and $\forall j \in C, s_j = (a_j, R_j, F, n_j)$, then:

   • If $k = \min\{\arg\max_{j \in C} n_j\}$ and $a^k \in \cup_{j \in C} L_j(a, R)$, then $g(s) = a^k$
   • Otherwise, $g(s) = a$

3) If $\forall i \in \mathbb{N}, s_i = (a^i, R^i, F, n^i)$, then $k = \min\{\arg\max_{j \in \mathbb{N}} n_j\}$ and set $g(s) = a^k$.

4) If none of the above apply, set $g(s) = a_H$.

Proof of Theorem 2

We will show that any SCR $f$ that satisfies our premises, namely UWA, WPD and WPO 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 $R^\ast$. Consider the strategy profile where $\forall i \in \mathbb{N}, s_i = (a, R^\ast, NF, \cdot)$. If $j \in \mathbb{N}$ deviates she will obtain any $b \in L_j(a, R^\ast)$. So, $g(S_j, s_{N\setminus\{j\}}) = L_j(a, R^\ast)$. If any $C \subset \mathbb{N}$ deviates, the obtained outcome will be in $L_j(a, R^\ast)$ for at least one $j \in C$. Finally, if $\mathbb{N}$ deviate, there cannot be an improvement for all $i \in \mathbb{N}$ since $f$ satisfies WPO. Therefore, $s$ is a strong equilibrium in $R^\ast$.

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

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

Claim 1: There is no strong equilibrium under rule 1 where $\forall i \in \mathbb{N}, R^i \neq R^\ast$.
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. \( \square \)

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 \( i \in N \setminus C, R^i \neq R^* \). \( \square \)

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 \( l = \min\{\text{argmax}_{j \in C} n_j^j\} \), we distinguish two cases:
  - \( l \neq i \): In this case, since \( a^l \in \bigcup_{j \in C} L_j(a, R^*) \), 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, \( a^l \in \bigcup_{j \in C} L_j(a, R^*) \), 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 \( \exists i \in C \), 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, \( h \), can deviate to \( s_h' = (b, R^*, F, n^h) \in T_i^G(R^*) \) and obtain \( b \) while announcing the true state \( R^* \), for a sufficiently high integer \( n^h \). Therefore, \( s \) cannot be a strong equilibrium.

**Corollary 2.** In any strong equilibrium \( s \) of the mechanism \( G \), it holds that \( \forall i \in N, R^i = R^* \).

*Proof.* Immediate implication of claims 1-4 as well as of the fact that there cannot exist any strong equilibria under rule 4.

By the above arguments, we can restrict attention to strong equilibria 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, \( \forall i \in N \setminus C, s_i = (a, R^*, F, n_i) \) 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 \( \forall i \in N \setminus C, L_i(a, R^*) \subseteq \cup_{j \in C} L_j(a, R^*) \subseteq L_i(b, R^*) \) 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 \( L_i(a, R^*) \subseteq L_i(b, R^*) \). 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 \( L_i(a, R^*) \subseteq L_i(b, R^*) \). 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 \( g(S_i, s_{N \setminus \{i\}}) = A \). Since \( f(R^*) \neq \emptyset \), there exists \( a \in f(R^*) \) such that \( \forall i \in N, L_i(a, R^*) \subseteq A \). Moreover, for \( s \) to be a strong equilibrium it must be that \( \forall i \in N, L_i(a, R^*) \subseteq A \subseteq L_i(b, R^*) \). Then, from WPD it must hold that \( b \in f(R^*) \).

This completes the proof.
References


