# SYMMETRIC MECHANISM DESIGN

YARON AZRIELI AND RITESH JAIN

Abstract. Designers of economic mechanisms often have an incentive to bias the rules of the mechanism in favor of certain groups of agents. This paper studies the extent to which a policy prohibiting biased mechanisms is effective in achieving fair outcomes. Our main result is a characterization of the class of social choice functions that can be implemented by symmetric mechanisms. When the solution concept used is Bayes-Nash equilibrium, symmetry is typically not very restrictive and discriminatory social choice functions can be implemented by symmetric mechanisms. Our characterization in this case is based on a 'revelation principle' type of result, where we show that a social choice function can be symmetrically implemented if and only if a particular kind of (indirect) symmetric mechanism implements it. When implementation in dominant strategies is considered, only symmetric social choice functions can be implemented by symmetric mechanisms. We illustrate our results in environments of voting with private values, voting with a common value, and assignment of indivisible goods.

Keywords: Mechanism design; symmetry; fairness; revelation principle.

JEL Classification: D82

April 14, 2016

Dept. of Economics, The Ohio State University, 1945 North High Street, Columbus, Ohio 43210, U.S.A. Email: azrieli.2@osu.edu and jain.224@osu.edu.

## 1. Introduction

Designers of economic mechanisms can often benefit by biasing the rules in favor of some of the participating agents. In auction design, for example, it is well known ([11]) that if bidders are heterogenous then the seller can intensify the competition by subsidizing the bids of weaker bidders. Thus, this 'affirmative action' type of bias may increase the seller's revenue.[1] Another example is the admission criteria employed by academic institutions in order to achieve a diverse student body, criteria that have been criticized as unfair and even challenged in courts.[2] In electoral competition, the practice of manipulating district boundaries to obtain partisan advantage (gerrymandering) often comes under fire for preventing fair elections.

On the other hand, fairness is a high priority goal for policy makers. This is evident from the many existing acts and regulations whose goal is to guarantee that markets are not biased against certain groups in the population.[3] In the context of mechanism design, a natural candidate for an anti-discriminatory regulation is that the "rules of the game" are the same for all participants. In other words, the mechanism should be symmetric across agents. Symmetry is a normatively appealing property that has been used extensively in the social choice and mechanism design literature. It appears for example in the classic result of May [10] on simple majority rule, in the theory of social welfare functions (e.g. [8, Chapter 22]), and in the theory of values of cooperative games (e.g. [14]), among many others.

The goal of this paper is to analyze the extent to which symmetric mechanisms guarantee fair outcomes. We consider an abstract mechanism design environment with incomplete information, allowing for both correlated types and interdependent values. The output of a mechanism specifies a public outcome and a private outcome for each agent. For instance, in a public good provision problem a mechanism determines whether the good is provided or not (the public outcome) and the transfer required from each of the agents (the private outcomes) as a function of the profile of messages sent to the mechanism.[4] Roughly speaking, a symmetric mechanism is one in which a permutation of

[1]In a recent paper, Deb and Pai [5] show how the seller can implement the revenue-maximizing auction without explicitly biasing the rules in favor of weaker bidders. We discuss their result and its relation to our work in detail below.

[2]A recent example is the case of Fisher verses University of Texas, US supreme court, vol. 570 (2013).

[3]For example, the U.S. Equal Employment Opportunity Commission (EEOC) enforces several anti-discriminatory labor-market laws. Another example is the Genetic Information Nondiscrimination Act of 2008 in health insurance markets.

[4]Environments with only public outcomes (such as voting environments) and environments with only private outcomes (such as auctions) can of course be accommodated in our framework.

the vector of messages results in no change to the public outcome and the corresponding permutation of the private outcomes.

Our main result (Theorem 1) characterizes the class of social choice functions (i.e., mappings from type profiles to outcomes) that a designer can implement in Bayes-Nash equilibrium using symmetric mechanisms. In many environments this class is large and contains functions that clearly favor certain agents over others. In some environments this includes even the extreme case of dictatorial functions, as we illustrate in Section 2 below. Thus, a regulation requiring mechanisms to treat agents symmetrically (in the sense we have defined it) is not necessarily an effective way to achieve fairness. On the other hand, in many cases symmetry does put some restrictions on what can be implemented, and combining symmetry with additional context-specific requirements may be an effective policy.

Our characterization can be roughly described as follows. A given (incentive compatible) social choice function $f$ can be implemented by a symmetric mechanism if and only if for every pair of agents $i$ and $j$ there exists another social choice function $f_{ij}$ such that (1) $f_{ij}$ treats $i$ and $j$ symmetrically; and (2) for any type of agent $i$, the expected utility $i$ receives under $f$ by truthfully revealing his type is weakly higher than the expected utility he receives under $f_{ij}$ from any possible type report. When such functions $f_{ij}$ exist, we explicitly show how to construct a symmetric mechanism that implements $f$. The construction is based on the idea that the equilibrium message of an agent encodes his identity as well as his type, and the mechanism uses $f$ to determine the outcome when the message profile contains all identities. The function $f_{ij}$ is used to incentivize agent $i$ to reveal his true identity instead of 'pretending' to be agent $j$.[5] Conversely, if for some pair of agents $i$ and $j$ an appropriate function $f_{ij}$ does not exist, then, not only that this construction does not work, symmetric implementation is impossible altogether. Thus, our result has the flavor of a 'revelation principle', in the sense that one should only consider a particular kind of (indirect) mechanisms to determine whether symmetric implementation is possible or not.

[5]This construction is reminiscent of mechanisms used in the literature on implementation with complete information (e.g. [9]), where each agent reports everyone's preferences and an agent is punished if his message disagrees with the messages of all other agents. There are important differences however: First, and most importantly, in our case the designer does not use the punishment scheme to elicit any new information from the agents; it is only used to create symmetry in cases where the original social choice function is not symmetric. Second, under complete information the construction works for *every* social choice function, while this is not the case in our setup. Third, our construction is meaningful even if there are only two agents, which is not the case with complete information.

We emphasize that we do not argue that the type of mechanism described above is a practical way for designers to get around a rule requiring symmetry. It is merely a theoretical construction that enables us to prove the sufficiency part of our theorem and to provide an upper bound on what can be symmetrically implemented. It may very well be that simpler and more practical symmetric mechanisms can be found that implement a given social choice function, as is the case in [5] for example. However, we do think that one of the contributions of this paper is to expose the role that indirect mechanisms have in overcoming exogenous constraints such as symmetry.[6]

A well-known criticism of implementation in Bayes-Nash equilibrium is that it assumes too much common knowledge among the agents and the planner (see [2]). One solution to this problem is to consider the more robust concept of dominant strategies. Dominant strategy mechanisms are typically studied in environments with private values, so we restrict attention to this type of environments. In Theorem 2 we prove that only symmetric social choice functions can be implemented in (weakly) dominant strategies by symmetric mechanisms. In other words, indirect mechanisms are not useful in overcoming the symmetry constraint. This stands in stark contrast to the results for implementation in Bayes-Nash equilibrium.

The closest paper to ours is Deb and Pai [5], who study the possibility of a seller to design discriminatory auctions in an independent private-values setup under the same symmetry constraint as in the current paper. However, the set of feasible mechanisms they consider is further restricted to include only mechanisms in which bidders submit a single number (the bid), and the highest bidder wins the object. This rules out the type of mechanisms that we use in our proof. Nevertheless, they are able to show that essentially any incentive compatible direct mechanism (social choice function in our terminology) can be implemented in Bayes-Nash equilibrium by some such symmetric auction.[7] In other words, symmetry puts almost no restriction on what can be implemented. In Corollary 1 below we show that the same is true in voting environments with independent private-values, i.e. *every* incentive compatible social choice function can be symmetrically implemented in Bayes-Nash equilibrium. However, as we show by examples, this is no longer true with correlated types or interdependent values.

[6]Other examples that illustrate the potential importance of indirect mechanisms include [3] for the case where the designer lacks commitment power, [12] when agents have menu-dependent preferences, and [13] when only deterministic mechanisms are allowed.

[7]The notion of implementation used in [5] is somewhat weaker than ours, since they only require that the *expected* payment in the indirect mechanism is equal to the payment in the direct mechanism. Our definition requires that they are equal ex-post.

In the following section we illustrate our result with three examples. The first is in an environment of voting with private values, the second in a common-value voting setup (as in the 'Condorcet Jury' theorem), and the third is in a model of assignment of indivisible goods. Section 3 presents the notation and the definition of symmetry that we use, and Section 4 contains the main result and its proof. The special case in which the outcome space contains only public outcomes (voting environment) is further analyzed in Section 5. Section 6 contains the analysis of implementation in dominant strategies. Section 7 concludes.

## 2. Motivating examples

2.1. **Voting with private values.** Consider a society of two agents which needs to choose an alternative from the set $\{x, y\}$ (they may also choose to randomize between the alternatives). Each agent can either be of type $X$ or of type $Y$. Both agents have the same (private-values) utility function given by $u(X, x) = u(Y, y) = 1$ and $u(X, y) = u(Y, x) = 0$, that is an agent of type $X$ prefers alternative $x$ and an agent of type $Y$ prefers alternative $y$. The realized profile of types is drawn according to the distribution given by the following matrix, where rows correspond to types of agent 1 and columns to types of agent 2:

| | $X$ | $Y$ |
|---|---|---|
| $X$ | $\mu$ | $1/2-\mu$ |
| $Y$ | $1/2-\mu$ | $\mu$ |

The parameter $0 \leq \mu \leq 1/2$ measures the likelihood that agents have similar preferences. Notice that when $\mu > 1/4$ agents' types are 'positively correlated' and when $\mu < 1/4$ they are 'negatively correlated'. When $\mu = 1/4$ types are independent.

Consider the case where agent 1 is a dictator who gets to choose his favorite alternative. We can describe this Social Choice Function (SCF) $f$ by the following matrix, where rows correspond to types of agent 1 and columns to types of agent 2:

| | $X$ | $Y$ |
|---|---|---|
| $X$ | $x$ | $x$ |
| $Y$ | $y$ | $y$ |

Clearly $f$ is incentive compatible, since agent 1 always gets his favorite alternative and agent 2 cannot influence the outcome. However, $f$ treats the agents asymmetrically, favoring agent 1 over 2. The violation of symmetry is expressed by the fact that

$f(X,Y)=x$ while $f(Y,X)=y$, so that two type profiles which are permutations of one another lead to different outcomes.

Is there a way to implement $f$ as a Bayesian equilibrium of some symmetric indirect mechanism? We claim that the answer is positive if and only if $\mu \geq 1/4$. To see this, consider the indirect mechanism in which the set of messages available to each agent is $M=\{m_1,m_2,m_3,m_4\}$, and the mapping from profiles of messages to alternatives is given by the following matrix:

| | $m_1$ | $m_2$ | $m_3$ | $m_4$ |
|---|---|---|---|---|
| $m_1$ | $\beta$ | $\gamma$ | $\boldsymbol{x}$ | $\boldsymbol{x}$ |
| $m_2$ | $\gamma$ | $\delta$ | $\boldsymbol{y}$ | $\boldsymbol{y}$ |
| $m_3$ | $x$ | $y$ | $x$ | $x$ |
| $m_4$ | $x$ | $y$ | $x$ | $x$ |

In the above matrix $\beta, \gamma$ and $\delta$ stand for numbers in $[0,1]$ representing the probability with which alternative $x$ is chosen ($y$ is chosen with the complementary probability). Notice that this is a symmetric mechanism in the sense that when the agents switch their messages the outcome does not change.

Consider the strategy profile in which agent 1 plays $\sigma_1(X)=m_1$ and $\sigma_1(Y)=m_2$ and agent 2 plays $\sigma_2(X)=m_3$ and $\sigma_2(Y)=m_4$. Note that if the agents follow these strategies then the SCF $f$ is implemented, as can be seen in the boldfaced entries in the matrix. Also, notice that $\sigma_1$ is a best response for agent 1 to $\sigma_2$. To make this profile of strategies an equilibrium we therefore need to choose $\beta,\gamma,\delta \in [0,1]$ so that the following inequalities are satisfied:

$$(2\mu)*\beta+(1-2\mu)*\gamma \leq 2\mu \tag{1}$$

$$(2\mu)*\gamma+(1-2\mu)*\delta \leq 2\mu \tag{2}$$

$$(1-2\mu)*(1-\beta)+(2\mu)*(1-\gamma) \leq 2\mu \tag{3}$$

$$(1-2\mu)*(1-\gamma)+(2\mu)*(1-\delta) \leq 2\mu \tag{4}$$

Inequalities (1) and (2) are the incentive constraints for agent 2 of type $X$ not to deviate to messages $m_1$ or $m_2$, respectively. Similarly, inequalities (3) and (4) are the incentive constraints for agent 2 of type $Y$.

Now, if $\mu \geq 1/4$ then it is immediate to check that $\beta=\gamma=\delta=1/2$ satisfy the above inequalities, and hence $f$ can be implemented by a symmetric mechanism. On the other hand, if $\mu < 1/4$ then a simple manipulation of (1) and (3) implies that $\beta=1$ and $\gamma=0$

must hold. But if $\gamma = 0$ then (4) cannot be satisfied, hence there is no solution to this system.

This argument only implies that the particular kind of mechanism considered above cannot symmetrically implement $f$ when $\mu < 1/4$. But our main result (Theorem 1) says that looking at this kind of mechanisms is *without loss of generality*, that is, it's impossible to find any symmetric mechanism that implements $f$. To understand this result better, let us rename the messages in the above mechanism so that $M = \{1, 2\} \times \{X, Y\}$. Thus, each agent reports a name of an agent and a type. Replace $m_1$ by $1X$, $m_2$ by $1Y$, $m_3$ by $2X$, and $m_4$ by $2Y$. The strategy profile we considered is the one in which each agent truthfully reports his name and type. The outcome function of the mechanism is designed so that under this truthful strategy profile the given SCF $f$ is used to determine the outcome. To make this truthful reporting an equilibrium, each agent should be incentivized not to deviate to messages that 'belong' to the other agent. It is easy to design a punishment for agent 1 to prevent him from 'pretending' to be agent 2, since agent 1 gets his favorite alternative in equilibrium. However, when $\mu < 1/4$ it is not possible to design an effective punishment for agent 2 to prevent him from 'pretending' to be 1. The difficulty comes from the fact that the punishment should satisfy the symmetry constraint (the outcome should be $\gamma$ in both message profiles $(1X, 1Y)$ and $(1Y, 1X)$). We show in Theorem 1 that if such a symmetric punishment scheme is not available then no symmetric mechanism that implements $f$ exists.

2.2. **Common value voting.** Consider a 'Condorcet Jury' setup, where given the state of nature all agents agree on the correct decision, but there is uncertainty about the realized state. Specifically, assume that the state is either $H$ or $L$, and that each of a group of $n$ agents receives a signal in $\{h, l\}$. Let $\mu$ be the (full-support) distribution over $\{H, L\} \times \{h, l\}^n$ according to which the state and signals are drawn. The set of possible decisions is $\{H, L\}$. All agents receive utility of 1 if the decision matches the state and a utility of 0 otherwise. To fit with our formal framework, it will be convenient to define the utility of agents as a function of the realized profile of signals and the decision chosen (by taking expectation over the state of nature). That is, the utility function of each agent $i$ is given by $u_i(\underline{t}, H) = \mu(H|\underline{t})$ and $u_i(\underline{t}, L) = \mu(L|\underline{t}) = 1 - u_i(\underline{t}, H)$.

Consider the SCF $f$ which chooses the decision more likely to be correct given any profile of types. That is $f(\underline{t}) = H$ if $\mu(H|\underline{t}) \geq 1/2$ and $f(\underline{t}) = L$ otherwise. Clearly $f$ is incentive compatible, since even if agent $i$ knows everyone else's types it is still a best response for him to be truthful (i.e., truth-telling is an ex-post equilibrium under $f$). However, since we make no assumption on the structure of $\mu$, it may be that some agents' signals are highly correlated with the state and hence provide valuable information, while others' signals are independent of the state. If this is the case then $f$ would not be symmetric – the decision will not depend just on the number of $h$ signals in the message profile but also on the identity of the agents who sent those $h$ signals.

It is possible however to implement $f$ by a symmetric mechanism. To see this, consider the 'names mechanism' described in the previous example, where the set of messages available for each agent is $M = \{1, \ldots, n\} \times \{h, l\}$. The mapping from message profiles to decisions is defined as follows. If all the names 1 through $n$ appear in the message profile then choose the decision prescribed by $f$ at the reported type profile, treating the names as if they were truthfully announced (even if they were not). In all other message profiles choose alternative $H$. This mechanism is clearly symmetric, in the sense that when the message profile is permuted the decision is not affected. In addition, the strategy profile in which each agent $i$ reports his real name and his type is an equilibrium, since in this case $f$ is used to determine the decision, so $i$ gets his preferred alternative at each type profile. Thus, this symmetric mechanism implements $f$.

2.3. **An assignment model.** Three goods $\{a, b, c\}$ are to be allocated to three agents $\{1, 2, 3\}$, one good per agent.[8] Each agent's type is drawn from the set

$$T = \{abc, acb, bac, bca, cab, cba\},$$

where each $t \in T$ corresponds to a strict ordering over the goods. Agents' types are independent and identically distributed according to the probabilities $\mu(abc) = \frac{1}{6} + 5\delta$ and $\mu(t) = \frac{1}{6} - \delta$ for any other $t \in T$, with $\delta \in [0, \frac{1}{6})$. Note that when $\delta = 0$ all types are equally likely, and that as $\delta$ increases the ordering $a \succ b \succ c$ becomes more likely relative to all other types. For concreteness, suppose that the utility an agent gets from consuming his most preferred good is 3, from consuming his second-best is 2, and from consuming his least favorable is 1.

A mechanism specifies a lottery over assignments as a function of the agents' messages. Our definition of symmetry requires in this setup that if the profile of messages is permuted then the vector of lotteries over the goods is permuted accordingly. For example, suppose that under some message profile agent 1 gets good $a$ for sure while agents 2 and 3 each gets good $b$ with probability 1/2 or good $c$ with probability 1/2.[9]

[8]This example is based on the assignment model of [7]. It is important for our argument that the designer must allocate all three goods and cannot leave some agents empty-handed, as is often the case in applications.

[9]By the Birkhoff–von Neumann Theorem this collection of lotteries corresponds to some distribution over assignments.

Symmetry implies that if agents 1 and 2 switch their messages (and agent's 3 message is unchanged) then agent 2 gets good $a$ for sure and agents 1 and 3 each gets good $b$ with probability 1/2 or good $c$ with probability 1/2. Note that this implies in particular that when two agents send the same message to the mechanism they should get the same lottery over goods.

Consider the 'serial dictatorship' SCF in which agent 1 gets to choose his favorite good first, and agent 2 gets to choose his favorite good from what's left after 1's choice. Agent 3 is assigned the good not chosen by the first two agents.[10] This SCF clearly does not satisfy the above symmetry requirement. Yet, is there a symmetric indirect mechanism that can implement serial dictatorship? It turns out that the answer depends on the parameter $\delta$; if $\delta$ is sufficiently small, specifically $\delta \leq \delta^* := \frac{-3+\sqrt{13}}{12} \cong 0.05$, then symmetric implementation is possible; if $\delta > \delta^*$ then it is not.

We present here a sketch of the argument for the above claim; a detailed proof appears in Appendix A. Consider again the 'names mechanism', and recall that the key for symmetric implementation is the existence of symmetric punishment schemes that prevent agents from sending other agents' messages. Since agent 1 always gets his favorite good according to the given SCF, it is easy to construct such a punishment scheme to prevent him from lying for any value of $\delta$. The same is true for agent 2, since if for example he pretends to be agent 3 then we can use the SCF in which agent 1 picks his favorite good as before, and the other two goods are randomly (uniformly) assigned to agents 2 and 3. This SCF is symmetric between agents 2 and 3, and agent 2 is worse off under this function relative to the original one for any value of $\delta$. A similar construction works when agent 2 pretends to be agent 1.

Symmetric implementation therefore depends on whether we can prevent agent 3 from pretending to be one of the other agents. The cutoff $\delta^*$ is obtained by equating the probability with which agent 3 gets item $c$ under the given serial dictatorship to 1/2: If $\delta \leq \delta^*$ then this probability is less or equal to 1/2, and if $\delta > \delta^*$ then it is greater than 1/2. If this probability is below 1/2 then the punishment scheme is simply the constant SCF in which the lottery over the goods for agent 3 and the agent he pretends to be are both equal to the lottery that 3 gets under the serial dictatorship. Since the probabilities in this lottery are all below 1/2 this yields a legitimate random assignment, and agent 3 has no incentive to deviate. However, if this probability is greater than 1/2 then this construction no longer works, and we show in Appendix A that nothing else works in

[10]See [1] for an analysis of the 'random serial dictatorship' mechanism in this model. In [4] this mechanism is referred to as 'random priority'.

this case. By our main result this implies that no symmetric mechanism can implement serial dictatorship.[11]

## 3. Notation and definitions

We will work in a standard abstract mechanism design environment with incomplete information as in e.g. [6]. There is a finite set $[n] = \{1, \ldots, n\}$ of agents, with $n \geq 2$. The private information of agents is described by their *types*, where $T$ is the finite set of possible types of each agent. A *state* or a *type profile* is an assignment of an element in $T$ for each agent, and we denote a state by $\underline{t} = (t_1, \ldots, t_n) \in T^n$, where $t_i \in T$ is the type of agent $i$. We use the standard notation in which a subscript $-i$ means that the $i^{th}$ coordinate of a vector is omitted. In particular, given a state $\underline{t}$, we write $t_{-i}$ for the vector of types of all agents except $i$, and we write $(t'_i, t_{-i})$ for the state in which the type of $i$ is $t'_i$ and the types of all other agents are the same as in $\underline{t}$.

The set of finite-support distributions over some set $A$ is denoted $\Delta(A)$.[12] Agents' beliefs about the types of others may depend on their own type in an arbitrary way. For each $i \in [n]$ there is a function $\mu_i : T \to \Delta\left(T^{n-1}\right)$ which specifies the beliefs of agent $i$ given each one of his possible types. For any state $\underline{t}$ we write $\mu_i(t_{-i}|t_i)$ for the probability that $i$ assigns to the other agents' type profile being $t_{-i}$ when his type is $t_i$. Note that we allow for correlated types and that beliefs need not be consistent with a common prior. If every $\mu_i$ is constant, that is $\mu_i(t_{-i}|t_i) = \mu_i(t_{-i}|t'_i)$ for all $i$, $t_i$, $t'_i$ and $t_{-i}$, then we say that types are *independent*.

Let $O$ be a set whose elements are called *public outcomes*, and let $P$ be a set whose elements are called *private outcomes*. For example, in a public good provision problem $O$ may stand for whether the good is provided or not and $P$ for the transfer required from the agent to finance the good. Agents' preferences are defined over $T^n \times O \times P$. Specifically, we assume that agent $i$'s preferences are represented by the von Neumann-Morgenstern utility function $u_i : T^n \times O \times P \to \mathbb{R}$. Note that this allows the possibility of interdependent values, where an agent preferences depend on other agents' types. Agents are assumed to be expected utility maximizers. With abuse of notation, if $\lambda \in \Delta(O \times P)$ then we write $u_i(\underline{t}, \lambda)$ for the expected utility $\sum_{o \in O} \sum_{p \in P} \lambda(o,p) u_i(\underline{t}, o, p)$.

[11] It is easy to generalize this example to any number of agents and any (i.i.d.) type distribution. The key to whether symmetric implementation is possible or not is the lottery over the goods for the last agent: If this agent receives one of the goods with probability greater than 1/2 then symmetric implementation is not possible, otherwise it is.

[12] Formally, $\Delta(A)$ is the set of all mappings $\varphi : A \to [0,1]$ such that $\varphi(a) \neq 0$ for only finitely many $a \in A$ and $\sum_{a \in A} \varphi(a) = 1$. We often identify each $a \in A$ with the lottery that assigns probability 1 to $a$.

The last component in the description of the environment is a set $D \subseteq O \times P^n$, whose elements are called *feasible outcomes*. Each $d = (o, \underline{p}) \in D$ is composed of a public outcome $o \in O$ and a vector $\underline{p} = (p_1, \ldots, p_n) \in P^n$ that specifies the private outcome for each agent. By allowing $D$ to be a strict subset of $O \times P^n$ we can incorporate constraints such as budget-balance in a public good provision problem, or that each agent gets a single good in an assignment problem (as in subsection 2.3 above). Throughout the paper we assume that $D$ is closed under permutations of the vector of private outcomes. Formally, if $\pi : [n] \to [n]$ is any permutation of the agents' names and $(o, \underline{p}) \in D$ then also $(o, \pi\underline{p}) \in D$, where $\pi\underline{p} = (p_{\pi(1)}, \ldots, p_{\pi(n)})$.

A *mechanism* specifies the set of *messages* available to each player and an *outcome function* which maps each profile of messages to a lottery over feasible outcomes. We will restrict our attention to mechanisms in which all players have the same finite set of messages. A mechanism is therefore formally defined as a pair $(M, g)$, where $M$ is some finite set and $g : M^n \to \Delta(D)$.

Before stating the main definition of this section we need one more piece of notation. Given any vector $\underline{x}$ of length $n$ and any pair $i, j \in [n]$, we denote by $\underline{x}^{ij}$ the vector of length $n$ which is identical to $x$ except that the $i$th and $j$th coordinates are switched.

**Definition 1.** A mechanism $(M, g)$ is *$ij$-symmetric* if for every $\underline{m} \in M^n$ and every $(o, \underline{p}) \in D$,

$$g(\underline{m})(o, \underline{p}) = g(\underline{m}^{ij})(o, \underline{p}^{ij}).$$

A mechanism $(M, g)$ is *symmetric* if it is $ij$-symmetric for every pair $i, j \in [n]$. Equivalently, $(M, g)$ is symmetric if for every permutation $\pi$ of the agents

$$g(\underline{m})(o, \underline{p}) = g(\pi\underline{m})(o, \pi\underline{p}).$$

In words, symmetric mechanisms have the property that if one message profile is a permutation of another then the mechanism output at the former is the corresponding permutation of the latter. For deterministic output this simply means that the public outcome is not affected by permutations of the message vector and the private outcomes are permuted accordingly. When the output is stochastic we require that the joint distribution over $D$ is permuted. Note that symmetry implies in particular that the marginal distribution over the public outcome $o$ is not affected by permutations of the message profile, and that if $\underline{m}$ is such that $m_i = m_j$ then the marginal of $g(\underline{m})$ on $O \times P_i$ is the same as its marginal on $O \times P_j$. In addition, note that symmetry not only means that agents are treated equally in the outcomes they receive themselves, but also that agents are symmetric in the influence they have on other agents' outcomes.

**Remark.** There are other possible ways to define symmetry that capture the idea that agents are treated equally by the mechanism. For instance, one could use a weaker notion than Definition 1 only requiring that the marginal on $O \times P_i$ of the distribution $g(\underline{m})$ is the same as the marginal on $O \times P_j$ of the distribution $g(\underline{m}^{ij})$, and that for any other agent $k$ the marginal on $O \times P_k$ is the same under these two message profiles. This definition makes sense since agents only care about the public outcome and their own private outcome. The two definitions coincide for deterministic mechanisms. Our results can be adapted to this alternative definition. However, we find Definition 1 to be the most natural one.

Given the beliefs and preferences of the players, every mechanism induces a (finite) Bayesian game. A (pure) strategy for an agent is a mapping $\sigma : T \to M$. If $\underline{\sigma} = (\sigma_1, \dots, \sigma_n)$ is a strategy profile and $\underline{t}$ is a state then we write $\underline{\sigma}(\underline{t}) = (\sigma_1(t_1), \dots, \sigma_n(t_n))$ for the vector of induced messages. Given some $\lambda \in \Delta(D)$, it will be convenient to denote by $\lambda_{O,i}$ the marginal of $\lambda$ on $O \times P_i$, that is $\lambda_{O,i}(o,p) = \sum_{\{\underline{p} \in P^n \,:\, p_i = p\}} \lambda(o, \underline{p})$ for every $(o,p) \in O \times P$. A strategy profile $\underline{\sigma}$ is a *Bayes-Nash Equilibrium (BNE)* of a mechanism $(M,g)$ if

$$\sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, g(\underline{\sigma}(\underline{t}))_{O,i}) \geq \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, g(m_i, \sigma_{-i}(t_{-i}))_{O,i})$$

for every $i \in [n]$, every $t_i \in T$, and every $m_i \in M$.

A *Social Choice Function (SCF)* is a mapping $f : T^n \to \Delta(D)$. The pair $(T, f)$ is itself a mechanism, called the *direct mechanism associated with* $f$. A SCF $f$ is $ij$-symmetric (symmetric) if the direct mechanism associated with $f$ is $ij$-symmetric (symmetric). *Truth-telling* refers to the strategy in the direct mechanism in which a player reports his true type. A SCF $f$ is *Bayesian Incentive Compatible (BIC)* if truth-telling is a BNE of the direct mechanism associated with $f$. Finally, we define the notion of symmetric implementation that will be used in the next sections.

**Definition 2.** Let $f$ be a SCF. A mechanism $(M,g)$ implements $f$ in BNE if there is a BNE $\underline{\sigma}$ of $(M,g)$ such that $f(\underline{t}) = g(\underline{\sigma}(\underline{t}))$ for every $\underline{t} \in T^n$.
We say that $f$ is *symmetrically implementable in BNE* if there is a symmetric mechanism $(M,g)$ that implements $f$ in BNE.

Note that we only require 'partial implementation', i.e. that there is an equilibrium for which the outcome of the mechanism coincides with that prescribed by the SCF. There may be other equilibria with different outcomes.

## 4. Main result

Before stating the result we need one more definition.

**Definition 3.** Fix an environment and two SCFs $f$ and $f'$. We say that $f$ *dominates* $f'$ *for agent* $i$ if

$$\sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, f(\underline{t})_{O,i}) \geq \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, f'(s_i, t_{-i})_{O,i}) \tag{5}$$

for every $t_i, s_i \in T$.

The left-hand side of (5) is the expected utility for agent $i$ of type $t_i$ under $f$ when all agents are truth-telling. The right-hand side is the expected utility for $i$ of type $t_i$ under $f'$ when he reports type $s_i$ and all other agents are truth-telling. Thus, $f$ dominates $f'$ for $i$ if the former is not smaller than the latter for any pair of types $t_i$ and $s_i$.

**Theorem 1.** A SCF $f$ is symmetrically implementable in BNE if and only if $f$ is BIC and for every ordered pair of agents $(i, j)$ there is an $ij$-symmetric SCF $f_{ij}$ such that $f$ dominates $f_{ij}$ for agent $i$.

*Proof.* We start with the 'If' direction. Assume that $f$ is BIC and that there are SCFs $\{f_{ij}\}$ satisfying the conditions in the statement of the theorem. We will now construct a symmetric mechanism $(M, g)$ and an equilibrium $\underline{\sigma}$ of this mechanism that implements $f$.

Let the set of messages available to each player be $M = [n] \times T$. Thus, each player reports to the mechanism a name of a player and a type, and we denote by $m_i = (b_i, s_i)$ the message sent by agent $i$, where $b_i \in [n]$ and $s_i \in T$. We describe the outcome function $g$ in three steps.

**Step 1:**
Consider message profiles in which each agent truthfully reports his name, that is for all $i$, $m_i = (i, s_i)$ for some $s_i \in T$. Define $g(\underline{m}) = g((1, s_1), \ldots, (n, s_n)) = f(s_1, \ldots, s_n)$ in this case.

Consider now some message profile $\underline{m} = ((b_1, s_1), \ldots, (b_n, s_n))$ in which $\{b_1, \ldots, b_n\} = [n]$. Any such $\underline{m}$ is a permutation of a unique message profile on which $g$ was just defined in the previous paragraph. Therefore, there is a unique way to extend $g$ to such message profiles while preserving symmetry. We define $g$ in this way. Note that our assumption that $D$ is closed under permutations of $\underline{p}$ guarantees that the range of $g$ is contained in $\Delta(D)$.

**Step 2:**
Fix two agents $i, j \in [n]$. Consider message profiles $\underline{m}$ in which $b_k = k$ for every $k \neq i$ and $b_i = j$, that is all agents truthfully report their name except agent $i$ who reports name $j$. Thus, the message profile is of the form

$$\underline{m} = ((1, s_1), \ldots, (i-1, s_{i-1}), (j, s_i), (i+1, s_{i+1}), \ldots, (j, s_j), \ldots, (n, s_n)).$$

Define $g(\underline{m}) = f_{ij}(s_1, \ldots, s_n)$, that is, the SCF $f_{ij}$ is used instead of $f$ on the reported types. Notice that the fact that $f_{ij}$ is $ij$-symmetric guarantees that our definition does not violate the symmetry of $g$ when $i$ and $j$ switch their messages.

Now, as in Step 1 above, there is a unique way to symmetrically extend $g$ to all message profiles in which $(b_1, \ldots, b_n)$ contains all names except $i$ and in which $j$ appears twice. We extend $g$ to preserve symmetry in this way. The same procedure is repeated for every ordered pair of agents $(i, j)$.

**Step 3:**
The message profiles on which $g$ has not been defined in the previous two steps are those in which at least two names are missing from the list $(b_1, \ldots, b_n)$. Define $g$ arbitrarily on such message profiles, while preserving symmetry. Given the assumption on $D$ such a symmetric function clearly exists.

The fact that the above mechanism is indeed symmetric is obvious from its construction. Consider the strategy profile defined by $\sigma_i(t_i) = (i, t_i)$ for every $i$, that is each player truthfully announces his name and his type. Clearly, $f(\underline{t}) = g(\underline{\sigma}(\underline{t}))$ for every $\underline{t} \in T^n$.

We claim that $\underline{\sigma}$ is a BNE of the mechanism. Indeed, note first that agent $i$ of type $t_i$ can't gain by deviating to a message of the form $(i, t'_i)$, since in this case the mechanism would still use $f$ to determine the outcome; since $f$ is BIC such misreporting of the type is not beneficial. Second, consider a deviation of agent $i$ of type $t_i$ to a message of the form $m_i = (j, s_i)$ with $j \neq i$. In this case the mechanism uses the function $f_{ij}$ to determine the outcome, and since $f$ dominates $f_{ij}$ for agent $i$ such a deviation is also not profitable. Thus, we have proven that $(M, g)$ symmetrically implements $f$ in BNE.

Consider now the 'Only If' direction, and assume that $f$ is symmetrically implementable in BNE. First, as is well-known, if $f$ is not BIC then it cannot be implemented in BNE by any mechanism, so in particular it cannot be implemented by a symmetric mechanism (the 'revelation principle'). Thus, $f$ must be BIC.

It is left to construct the collection of SCFs $\{f_{ij}\}$ from the theorem. Let $g$ be a symmetric mechanism that together with the equilibrium $\underline{\sigma}$ implements $f$. Given a pair of players $(i, j)$, define the strategy profile $\underline{\sigma}'$ by $\sigma'_k = \sigma_k$ for every $k \neq i$ and $\sigma'_i = \sigma_j$. That is, in $\underline{\sigma}'$ all players except $i$ play their equilibrium strategies, and player $i$ plays $j$'s equilibrium strategy. Define $f_{ij}(\underline{t}) = g(\underline{\sigma}'(\underline{t}))$.

We claim that this $f_{ij}$ satisfies the conditions in the theorem. First, $f_{ij}$ is $ij$-symmetric, since

$$f_{ij}(\underline{t})(o, \underline{p}) = g(\underline{\sigma}'(\underline{t}))(o, \underline{p}) = g((\underline{\sigma}'(\underline{t}))^{ij})(o, \underline{p}^{ij}) = g(\underline{\sigma}'(\underline{t}^{ij}))(o, \underline{p}^{ij}) = f_{ij}(\underline{t}^{ij})(o, \underline{p}^{ij}),$$

where the first and last equalities are by the definition of $f_{ij}$, the second equality follows from the symmetry of $g$, and the third equality follows from the fact that $i$ and $j$ play the same strategy under $\underline{\sigma}'$.

Second, for every two types $t_i, s_i \in T$ we have that

$$\begin{aligned}\sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, f(\underline{t})_{O,i}) \;=\; & \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, g(\underline{\sigma}(\underline{t}))_{O,i}) \geq \\ & \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, g(\sigma_j(s_i), \sigma_{-i}(t_{-i}))_{O,i}) = \\ & \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, g(\underline{\sigma}'(s_i, t_{-i}))_{O,i}) = \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(\underline{t}, f_{ij}(s_i, t_{-i})_{O,i}),\end{aligned}$$

where the first equality follows from $f(\underline{t}) = g(\underline{\sigma}(\underline{t}))$, the inequality from the fact that $\underline{\sigma}$ is a BNE, the next equality by the definition of $\underline{\sigma}'$, and the last equality by the definition of $f_{ij}$. Thus, $f$ dominates $f_{ij}$ for agent $i$ and the proof is complete. □

## 5. Voting environments

We call an environment a *voting environment* if it contains only public outcomes. This can be formally embedded in our general framework by letting $P$ be a singleton. For ease of notation we will identify $D$ with $O$ in this case. Agents' preferences are described by $u_i : T^n \times O \to \mathbb{R}$, and an outcome function of a mechanism is $g : M^n \to \Delta(O)$. Symmetry here means that $g$ is invariant under permutations of the vector of messages, that is $g(\underline{m}) = g(\pi\underline{m})$ for every permutation $\pi$.

### 5.1. **Voting with private values.**

We start by analyzing voting environments in which the preferences of a player do not depend on the types of his opponents.

**Definition 4.** Agents in a voting environment have *private values* if $u_i(\underline{t}, o) = u_i(\underline{t}', o)$ for every $i \in [n]$, every $o \in O$, and every $\underline{t}, \underline{t}' \in T^n$ with $t_i = t'_i$. When agents have private values we will write $u_i(t_i, o)$ for the utility of agent $i$ of type $t_i$ when alternative $o$ is chosen.

We now use Theorem 1 to derive two conditions, one necessary and one sufficient, for a SCF to be symmetrically implementable in voting environments with private values. These conditions are easier to check and to understand than the condition of Theorem 1, as will be illustrated by examples below. The following notation will be useful. Let

$$\lambda_i(t_i; f) = \sum_{t_{-i}} \mu_i(t_{-i}|t_i) f(t_i, t_{-i}) \in \Delta(O)$$

be the expected distribution over $O$ under a SCf $f$ when agent $i$ is of type $t_i$. Also, let

$$LC_i(t_i; f) = \{\lambda \in \Delta(O) \ : \ u_i(t_i, \lambda) \leq u_i(t_i, \lambda_i(t_i; f))\}$$

be the lower contour set in $\Delta(O)$ of the lottery $\lambda_i(t_i; f)$ for type $t_i$ of agent $i$, i.e. the set of lotteries which are weakly worse for $i$ of type $t_i$ than the lottery he can get by truthfully reporting his type.

**Proposition 1.** Consider a voting environment with private values.
1. Let $f$ be a BIC SCF. If $\bigcap_{t_i \in T} LC_i(t_i; f) \neq \emptyset$ for each agent $i$, then $f$ is symmetrically implementable in BNE.
2. If $n = 2$ and $f$ is symmetrically implementable in BNE, then $LC_i(t_i; f) \cap LC_i(t_i'; f) \neq \emptyset$ for each agent $i$ and for each pair of types $t_i, t_i' \in T$.

*Proof.* 1. Fix an agent $i$ and let $\lambda^* \in \bigcap_{t_i \in T} LC_i(t_i; f)$. For each other agent $j$ and for every $\underline{t}$ define $f_{ij}(\underline{t}) = \lambda^*$. Clearly, $f_{ij}(\underline{t}) = f_{ij}(\underline{t}^{ij})$ for all $\underline{t}$, so $f_{ij}$ is $ij$-symmetric. Also, for every $t_i, s_i \in T$ we have

$$\sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(t_i, f(\underline{t})) = u_i(t_i, \lambda_i(t_i; f)) \geq u_i(t_i, \lambda^*) = \sum_{t_{-i}} \mu_i(t_{-i}|t_i) u_i(t_i, f_{ij}(s_i, t_{-i})),$$

so $f$ dominates $f_{ij}$ for agent $i$. It follows from Theorem 1 that $f$ can be symmetrically implemented.

2. We will prove the result for agent $i = 1$, the proof for $i = 2$ is identical. Fix two types $t_1, t_1' \in T$. Since $f$ is symmetrically implementable in BNE, it follows from Theorem 1 that there is a symmetric function $f_{12} : T^2 \to \Delta(O)$ such that for every $s \in T$

$$u_1(t_1, \lambda_1(t_1; f)) \geq \sum_{t \in T} \mu_1(t|t_1) u_1(t_1, f_{12}(s, t)),$$

and

$$u_1(t_1', \lambda_1(t_1'; f)) \geq \sum_{t \in T} \mu_1(t|t_1') u_1(t_1', f_{12}(s, t)).$$

For each $s \in T$ multiply the first inequality by $\mu_1(s|t'_1)$ and sum up over all $s \in T$. Similarly, multiply the second inequality by $\mu_1(s|t_1)$ and sum up over all $s \in T$. We get

$$u_1(t_1, \lambda_1(t_1; f)) \geq \sum_{(s,t)\in T^2} \mu_1(s|t'_1)\mu_1(t|t_1)u_1(t_1, f_{12}(s,t)) = u_1\left(t_1, \sum_{(s,t)\in T^2} \mu_1(s|t'_1)\mu_1(t|t_1)f_{12}(s,t)\right),$$

and

$$u_1(t'_1, \lambda_1(t'_1; f)) \geq \sum_{(s,t)\in T^2} \mu_1(s|t_1)\mu_1(t|t'_1)u_1(t'_1, f_{12}(s,t)) = u_1\left(t_1, \sum_{(s,t)\in T^2} \mu_1(s|t_1)\mu_1(t|t'_1)f_{12}(s,t)\right).$$

But since $f_{12}(s,t) = f_{12}(t,s)$ for every pair $(s,t)$, it follows that

$$\sum_{(s,t)\in T^2} \mu_1(s|t'_1)\mu_1(t|t_1)f_{12}(s,t) = \sum_{(s,t)\in T^2} \mu_1(s|t_1)\mu_1(t|t'_1)f_{12}(s,t).$$

Hence, this lottery belongs to both $LC_i(t_i; f)$ and $LC_i(t'_i; f)$. □

**Remark.** In the simple case where $n = 2$ and $|O| = 2$ the sufficient condition of part 1 and the necessary condition of part 2 in Proposition 1 coincide, hence we get a characterization. Indeed, when $|O| = 2$ the set $\Delta(O)$ is one dimensional, so non-empty intersection of each pair of sets $LC_i(t_i; f)$, $LC_i(t'_i; f)$ implies that the whole collection intersects.

**Example 1.** We first illustrate the conditions in Proposition 1 by revisiting the example of subsection 2.1. Recall that $f$ is a dictatorial SCF in which agent 1 gets to choose his favorite alternative out of $\{x, y\}$. We identify $\Delta(O)$ with the $[0, 1]$ interval, so that any $p \in [0, 1]$ corresponds to the probability of $x$ being chosen.

If agent 1 is of type $X$ then $x$ is chosen for sure, and when he is of type $Y$ then $y$ is chosen for sure. In the above notation this means that $\lambda_1(X; f) = 1$ and $\lambda_1(Y; f) = 0$, which implies $LC_1(X; f) = [0, 1]$ and $LC_1(Y; f) = [0, 1]$. Thus, the condition in Proposition 1 holds for this agent for any value of $\alpha$.

For agent 2, $\lambda_2(X; f) = 2\alpha$ and $\lambda_2(Y; f) = 1 - 2\alpha$, so $LC_2(X; f) = [0, 2\alpha]$ and $LC_2(Y; f) = [1 - 2\alpha, 1]$. Thus, $LC_2(X; f) \cap LC_2(Y; f) \neq \emptyset$ if and only if $\alpha \geq 1/4$. Proposition 1 then tells us that $f$ can be symmetrically implemented in BNE if and only if $\alpha \geq 1/4$, as we saw in subsection 2.1.

**Example 2.** Let $n = 2$, $O = \{x, y, z\}$, and $T = \{X, Y, Z\}$. Both agents have the same utility function given by $u(X, x) = u(Y, y) = u(Z, z) = 1$ and $u = 0$ otherwise. Thus, each agent's expected utility is the expected probability of the alternative corresponding to his type being chosen. Agents' beliefs are the posteriors generated from the (common) prior $\mu$ given by the matrix

| | $X$ | $Y$ | $Z$ |
|---|---|---|---|
| $X$ | 0 | 1/3 | 0 |
| $Y$ | 0 | 0 | 1/3 |
| $Z$ | 1/3 | 0 | 0 |

We identify each distribution in $\Delta(O)$ with the vector of probabilities $(p_x, p_y, p_z)$. Let $\alpha = (4/9, 4/9, 1/9)$, $\beta = (1/9, 4/9, 4/9)$, and $\gamma = (4/9, 1/9, 4/9)$. Consider the SCF $f$ defined by the matrix

| | $X$ | $Y$ | $Z$ |
|---|---|---|---|
| $X$ | $\alpha$ | $\beta$ | $\gamma$ |
| $Y$ | $\alpha$ | $\beta$ | $\gamma$ |
| $Z$ | $\alpha$ | $\beta$ | $\gamma$ |

Note that $f$ is BIC, since agent 1's type does not affect the outcome at all, and agent 2 gets one of his favorite outcomes (out of $\alpha, \beta, \gamma$) for each of his types.

We now check the conditions of Proposition 1, starting with agent 2. We have $\lambda_2(X; f) = \alpha$, $\lambda_2(Y; f) = \beta$, and $\lambda_2(Z; f) = \gamma$. Thus, at every type, agent 2 has expected utility of 4/9 when he reveals his true type. It follows that the center of the simplex, i.e. the distribution $(1/3, 1/3, 1/3)$, is in the intersection of the lower contour sets for the three types, since for each type it gives agent 2 an expected utility of just 1/3.

Moving to agent 1, we have $\lambda_1(X; f) = \beta$, $\lambda_1(Y; f) = \gamma$, and $\lambda_1(Z; f) = \alpha$. Thus, agent's 1 expected utility when he truthfully reports his type is only 1/9. Since for any distribution in $\Delta(O)$ at least one outcome has probability greater than 1/9, it follows that the intersection of the three lower contour sets for agent 1 is empty. However, notice that for every pair of types of agent 1 the lower contour sets for these types intersect. For example, the distribution $(0, 0, 1)$ is in both $LC_1(X; f)$ and $LC_1(Y; f)$.

We conclude that the necessary condition of Proposition 1 is satisfied in this example, but the sufficient condition is not. We therefore cannot rely on this proposition to determine whether $f$ is symmetrically implementable or not, and instead we use directly the characterization in Theorem 1 to show that in fact the answer is positive. Indeed, we can define $f_{21}(\underline{t}) = (1/3, 1/3, 1/3)$ for every $\underline{t}$ as the 'punishment' SCF for agent 2 when he pretends to be agent 1. For agent 1, the punishment SCF $f_{12}$ can be for example

| | $X$ | $Y$ | $Z$ |
|---|---|---|---|
| $X$ | $x$ | $y$ | $x$ |
| $Y$ | $y$ | $y$ | $z$ |
| $Z$ | $x$ | $z$ | $z$ |

It is straightforward to verify that these $f_{21}$ and $f_{12}$ satisfy the conditions of Theorem 1.

Finally, it is worth noting that the example can be modified so that the prior has full support. All the relevant inequalities in the example are strict, so for any prior sufficiently close to $\mu$ none of the arguments will be affected.

An important corollary of Proposition 1 is that, in voting environments with private values and independent types, symmetry constraints imposed on the mechanism never bind.

**Corollary 1.** In a voting environment with private values and independent types a SCF $f$ is symmetrically implementable in BNE if and only if $f$ is BIC.

*Proof.* We prove just the 'if' direction, as the 'only if' direction is clear. Let $f$ be some BIC SCF, and fix an agent $i$. Since $i$'s belief does not depend on his type, we simply write $\mu_i(t_{-i})$ instead of $\mu_i(t_{-i}|t_i)$. Fix some type $t_i^* \in T$ and denote $\lambda^* = \lambda_i(t_i^*; f)$. We claim that $\lambda^* \in \bigcap_{t_i \in T} LC_i(t_i; f)$ and hence that this intersection is non-empty. Indeed, for every $t_i \in T$,

$$u_i(t_i, \lambda^*) = \sum_{t_{-i}} \mu(t_{-i}) u_i(t_i, f(t_i^*, t_{-i})) \leq \sum_{t_{-i}} \mu(t_{-i}) u_i(t_i, f(t_i, t_{-i})) = u_i(t_i, \lambda_i(t_i; f)),$$

where the first equality is by the definition of $\lambda^*$, the inequality follows from the assumption that $f$ is BIC, and the last equality is by the definition of $\lambda_i(t_i; f)$. □

We point out that both assumptions of private values and independent types are needed for the result. The fact that not every BIC SCF can be implemented in voting environments with private values can be seen in Example 1 above. In Appendix B we provide an example of an environment with independent types but interdependent values and a BIC SCF in this environment that cannot be symmetrically implemented in BNE.

5.2. **Common value voting.** We now show how the example of subsection 2.2 generalizes to arbitrary common-value voting environments.

**Definition 5.** A *common value* voting environment is a voting environment in which $u_i(\underline{t}, o) = u_j(\underline{t}, o)$ for every $i, j \in [n]$, every $o \in O$, and every $\underline{t} \in T^n$. We denote by $u$ the common utility function of the agents.

**Proposition 2.** If a SCF $f$ in a common value voting environment satisfies $u(\underline{t}, f(\underline{t})) \geq u(\underline{t}, o)$ for every $o \in O$ and every $\underline{t} \in T^n$ (i.e., $f$ is utilitarian) then $f$ is symmetrically implementable in BNE.

*Proof.* Since $f$ is utilitarian, it follows that $u(\underline{t}, f(\underline{t})) \geq u(\underline{t}, f(t'_i, t_{-i}))$ for every $\underline{t}$, every $i$, and every $t'_i$. Thus, truth-telling is an ex-post equilibrium under $f$, so in particular $f$ is BIC. Choose an arbitrary outcome $o^* \in O$ and define $f_{ij}(\underline{t}) = o^*$ for every $i, j \in [n]$ and every $\underline{t} \in T^n$. Then clearly each $f_{ij}$ is symmetric, and since $f$ is utilitarian we have that $u(\underline{t}, f(\underline{t})) \geq u(\underline{t}, o^*) = u(\underline{t}, f_{ij}(s_i, t_{-i}))$ for every $\underline{t}$ and $s_i$. It follows that $f$ dominates $f_{ij}$ for agent $i$, so by Theorem 1 $f$ is symmetrically implementable in BNE. □

## 6. Symmetric implementation in dominant strategies

In this section we demonstrate that the possibility to use indirect mechanisms in order to symmetrically implement asymmetric SCFs is very limited when one requires implementation in dominant strategies rather than BNE. We focus here on environments with private values, since dominant strategy mechanisms are typically studied in such environments. Thus, the utility function of each agent $i$ depends only on his own type, and we view $u_i$ as a mapping $u_i : T \times O \times P \to \mathbb{R}$. In addition, to make our point particularly clear, we assume that the labeling of types is the same for all agents in the sense that $u_i(t, \cdot, \cdot) \equiv u_j(t, \cdot, \cdot)$ for any type $t \in T$ and every $i, j \in [n]$. This assumption is satisfied in the voting examples considered in the previous section, as well as in the assignment model of subsection 2.3.

Given a mechanism $(M, g)$, a strategy $\sigma_i$ of player $i$ is (weakly) *dominant* if for every $t_i \in T$ and every $m_i \in M$

$$u_i(t_i, g(\sigma_i(t_i), m_{-i})_{O,i}) \geq u_i(t_i, g(m_i, m_{-i})_{O,i})$$

holds for every $m_{-i} \in M^{n-1}$ with a strict inequality for at least one $m_{-i}$. A strategy profile $\underline{\sigma}$ is a *Dominant Strategy Equilibrium (DSE)* if $\sigma_i$ is a dominant strategy for all $i \in [n]$.

**Definition 6.** Let $f$ be a SCF. A mechanism $(M, g)$ implements $f$ in DSE if there is a DSE $\underline{\sigma}$ of $(M, g)$ such that $f(\underline{t}) = g(\underline{\sigma}(\underline{t}))$ for every $\underline{t} \in T^n$.
We say that $f$ is *symmetrically implementable in DSE* if there is a symmetric mechanism $(M, g)$ that implements $f$ in DSE.

**Theorem 2.** Under the assumptions of this section, if a SCF $f$ is symmetrically implementable in DSE then the direct mechanism associated with $f$ is symmetric.

In words, under the assumptions of this section, the use of indirect mechanisms does not increase the class of SCFs that can be symmetrically implemented in dominant strategies. Only SCFs that inherently treat the agents symmetrically can be implemented. This stands in stark contrast to the situation with BNE implementation, where indirect mechanisms can often be used to create artificial symmetry, as we saw in the previous sections.

*Proof.* Suppose that $f$ is symmetrically implementable in DSE. Let $(M, g)$ be a symmetric mechanism that together with the strategy profile $\underline{\sigma}$ implements $f$ in DSE. We will now show that it must be the case that $\sigma_1 = \ldots = \sigma_n$, i.e., all players use the same strategy. Notice that this will complete the proof, since it implies (together with the symmetry of $g$) that for any permutation $\pi$ of the agents, any $\underline{t} \in T^n$, and any $(o, \underline{p}) \in D$

$$f(\underline{t})(o, \underline{p}) = g(\underline{\sigma}(\underline{t}))(o, \underline{p}) = g(\pi(\underline{\sigma}(\underline{t})))(o, \pi\underline{p}) = g(\underline{\sigma}(\pi\underline{t}))(o, \pi\underline{p}) = f(\pi\underline{t})(o, \pi\underline{p}).$$

To show that all players use the same strategy, fix $i, j \in [n]$ and a type $t_* \in T$. Denote $m_* = \sigma_i(t_*)$. We need to show that $\sigma_j(t_*) = m_*$. Fix an arbitrary message $m \in M$ and a profile of messages $m_{-j}$ for the players other than $j$. Let $\underline{m}_* = (m_*, m_{-j})$ and $\underline{m} = (m, m_{-j})$. Notice that in the profile $\underline{m}_*^{ij}$ agent $i$ plays $m_*$, while in the profile $\underline{m}^{ij}$ agent $i$ plays $m$; all agents except $i$ send the same message in these two profiles. Thus, we get that

$$u_j(t_*, g(\underline{m}_*)_{O,j}) = u_i(t_*, g(\underline{m}_*^{ij})_{O,i}) \geq u_i(t_*, g(\underline{m}^{ij})_{O,i}) = u_j(t_*, g(\underline{m})_{O,j}),$$

where the two equalities follow from the symmetry of $g$ and our assumption on the labeling of types, and the inequality follows from the assumption that $m_* = \sigma_i(t_*)$ is a dominant message for $i$ of type $t_*$. It follows that $m_*$ is a best response for agent $j$ of type $t_*$ against any profile $m_{-j}$. Since there cannot be two weakly dominant strategies this implies that $\sigma_j(t_*) = m_*$ as needed. □

## 7. Final remarks

We have analyzed the extent to which a symmetry constraint limits the set of SCFs that a designer can implement. The main takeaway from the analysis is that symmetry by itself is typically not very restrictive if one considers implementation in Bayes-Nash equilibrium. In particular, symmetry does not prevent the implementation of highly biased SCFs. This apparent contradiction between the symmetry of the mechanism and the unfair outcomes it generates may seem surprising at first look, but notice that the bias is generated by the asymmetry in the equilibrium strategies of the players and not by the mechanism itself. When considering implementation in dominant strategies, the symmetry of the mechanism implies that all agents use the same strategy and hence that the resulting SCF is symmetric.

In some cases the agents participating in the mechanism have inherently different roles, for example buyers and sellers, or firms and workers. In such situations it seems more suitable to define symmetry only within each group of agents having the same role and not across all agents. While we do not pursue this direction here, we believe that such an extension would not create significant complications.

It is important to point out again that Theorem 1 concerns partial implementation, and that the mechanisms we consider often have multiple equilibria that may generate very different outcomes than the one we focus on. This is especially clear by looking at the example of subsection 2.1: If the agents switch their strategies, that is agent 1 plays $\sigma_1(X) = m_3$ and $\sigma_1(Y) = m_4$ and agent 2 plays $\sigma_2(X) = m_1$ and $\sigma_2(Y) = m_2$ then we get an equilibrium that implements the SCF in which agent 2 is the dictator instead of agent 1! This phenomenon is not unique to this particular example; similar multiplicity of equilibria will hold whenever the underlying environment is sufficiently symmetric. We leave a detailed study of full implementation under symmetry for future research.

Finally, given our results, an important question is what other regulations, either in isolation or combined with symmetry, are more effective in preventing implementation of discriminatory SCFs. Our analysis suggests that limiting the messages' space of the mechanism, either by imposing bounds on its cardinality or by prohibiting certain kinds of labeling of messages, may help in achieving this goal. Another possibility is to restrict the set of feasible outcomes ($D$ in our notation) so that it is more difficult for the designer to punish agents who deviate from their designated behavior. Exploring these interesting directions is beyond the scope of this paper.

## References


[1] A. Abdulkadiroğlu and T. Sönmez (1998), Random serial dictatorship and the core from random endowments in house allocation problems, *Econometrica* **66**, 689-701.

[2] D. Bergemann and S. Morris (2005), Robust mechanism design, *Econometrica* **73**, 1771-1813.

[3] H. Bester and R. Strausz (2000), Imperfect commitment and the revelation principle: the multi agent case, *Economics Letters* **69**, 165-171

[4] A. Bogomolnaia and H. Moulin (2001), A new solution to the random assignment problem, *Journal of Economic Theory* **100**, 295-328.

[5] R. Deb and M. Pai (2015), Discrimination via symmetric auctions, working paper.

[6] B. Holmström and R.B. Myerson (1983), Efficient and durable decision rules with incomplete information, *Econometrica* **51**, 1799-1819.

[7] A. Hylland and R. Zeckhauser (1979), The efficient allocation of individuals to positions, *The Journal of Political Economy* **87**, 293-314.

[8] A. Mas-Colell, M.D. Whinston, and J.R. Green (1995), *Microeconomic Theory*, Oxford University Press, New York.

[9] E. Maskin (1999), Nash equilibrium and welfare optimiality, *Review of Economic Studies* **66**, 23-38

[10] K.O. May (1952), A set of independent necessary and sufficient conditions for simple majority decisions, *Econometrica* **20**, 680-684.

[11] R.B. Myerson (1981), Optimal auction design. *Mathematics of operations research* **6**, 58-73.

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

[13] R. Strausz (2003), Deterministic mechanisms and the revelation principle, *Economics Letters* **79**, 333-337.

[14] L.S. Shapley (1953), A value for n-person games. In *Contributions to the Theory of Games II (Annals of Mathematical Studies 28)*, H.W. Kuhn and A.W. Tucker, editors. Princeton University Press.

## Appendix A. Proof of claim in the assignment example

**Claim:** In the assignment example described in subsection 2.3, serial dictatorship is symmetrically implementable in BNE if and only if $\delta \leq \delta^* := \frac{-3+\sqrt{13}}{12} \cong 0.05$.

*Proof.* We use the characterization in Theorem 1 to prove the claim. Denote by $f$ the given serial dictatorship SCF. First, $f$ is clearly BIC for any value of $\delta$. Second, if we define $f_{12}$ and $f_{13}$ to be the uniform random assignment of goods to agents then these functions satisfy the conditions in the theorem for any value of $\delta$. Indeed, these functions are clearly symmetric, and since agent 1 always gets his favorite good under $f$ his expected utility cannot increase under any alternative SCF and any possible report of type.

Next, define the SCF $f_{23}$ to be the one in which agent 1 gets his favorite good and agents 2 and 3 are equally likely to get each of the remaining goods. Then $f_{23}$ is 23-symmetric and is dominated by $f$ for agent 2. Indeed, at any realization of types, agent 2 gets his favorite good after 1's favorite is removed under $f$, while under $f_{23}$ he is equally likely to get either that good or a good he likes less, regardless of the type he reports. Similarly, define $f_{21}$ to be the SCF in which agent 3 gets his favorite good, and agents 1 and 2 are equally likely to get each of the other goods. By the symmetry between the agents, a similar argument shows that this SCF satisfies the conditions of the theorem as well.

It remains to check whether we can find appropriate SCFs $f_{31}$ and $f_{32}$ to prevent agent 3 from pretending to be one of the other agents. We claim that such functions exist if and only if $\delta \leq \delta^*$. Let $(p_a(\delta), p_b(\delta), p_c(\delta))$ be the probabilities with which agent 3 gets each of the goods under $f$ for a particular value of $\delta$. Notice that these probabilities do not depend on 3's type. A simple computation gives that

$$\begin{aligned} p_a(\delta) &= \frac{1}{3} - 4\delta + 12\delta^2, \\ p_b(\delta) &= \frac{1}{3} + \delta - 18\delta^2, \\ p_c(\delta) &= \frac{1}{3} + 3\delta + 6\delta^2, \end{aligned}$$

which implies that $p_a(\delta)$, $p_b(\delta)$ are decreasing and $p_c(\delta)$ is increasing for $\delta \in [0, 1/6)$.

We now argue that appropriate $f_{31}$ and $f_{32}$ exist if and only if $p_c(\delta) \leq 1/2$, which is equivalent to $\delta \leq \delta^*$ (note that $p_a(\delta)$ and $p_b(\delta)$ are always less than $1/2$). Suppose first that this condition holds. Define $f_{31}$ to be the constant SCF that at every type profile selects the random assignment given by the following matrix

| | $a$ | $b$ | $c$ |
|---|---|---|---|
| 1 | $p_a(\delta)$ | $p_b(\delta)$ | $p_c(\delta)$ |
| 2 | $1 - 2p_a(\delta)$ | $1 - 2p_b(\delta)$ | $1 - 2p_c(\delta)$ |
| 3 | $p_a(\delta)$ | $p_b(\delta)$ | $p_c(\delta)$ |

Since all three probabilities $p_a(\delta)$, $p_b(\delta)$ and $p_c(\delta)$ are in $[0, 1/2]$ this is a bi-stochastic matrix, and hence corresponds to some lottery over assignments. Since this SCF is constant and gives identical lotteries over the goods to agents 1 and 3 it is 31-symmetric as needed. Finally, agent 3 gets the same lottery over goods under $f$ and under $f_{31}$ regardless of his report, so inequality (5) holds as equality. The construction of the SCF $f_{32}$ is similar. We therefore proved that $f$ is symmetrically implementable in BNE when $\delta \leq \delta^*$.

Suppose now that $\delta > \delta^*$, so that $p_c(\delta) > 1/2$. Assume by contradiction that a 31-symmetric SCF $f_{31}$ exists that is dominated by $f$. For some fixed type $t \in T$, denote by $(q_a(\delta), q_b(\delta), q_c(\delta))$ the expected lottery over the goods that agent 3 obtains by reporting type $t$ under $f_{31}$. Then (5) implies that the following inequalities must hold:

$$\begin{aligned} 3p_a(\delta) + 2p_b(\delta) + p_c(\delta) &\geq 3q_a(\delta) + 2q_b(\delta) + q_c(\delta), \\ p_a(\delta) + 2p_b(\delta) + 3p_c(\delta) &\geq q_a(\delta) + 2q_b(\delta) + 3q_c(\delta). \end{aligned}$$

The first inequality means that if agent 3 is of type $abc$ then truthfully reporting his name (and his type) is not worse for him then pretending to be agent 1 of type $t$. The second inequality corresponds to agent 3 of type $cba$. Together, and taking into account that these are distributions, these inequalities imply that

$$p_a(\delta) - q_a(\delta) = p_c(\delta) - q_c(\delta).$$

Similarly, using the inequalities for types $acb$ and $bca$ of agent 3, we get that

$$p_a(\delta) - q_a(\delta) = p_b(\delta) - q_b(\delta)$$

must hold. Altogether this implies that $(p_a(\delta), p_b(\delta), p_c(\delta)) = (q_a(\delta), q_b(\delta), q_c(\delta))$.

Hence, since $t$ was arbitrary, under $f_{31}$ the expected lottery over the goods for agent 3 is the same regardless of his type, and is equal to the lottery he obtains under the serial dictatorship $f$. This implies in turn that under $f_{31}$ the ex-ante (before learning his type) expected lottery over the goods for agent 3 is also $(p_a(\delta), p_b(\delta), p_c(\delta))$. Since $f_{31}$ is 31-symmetric, and since the distribution of types is i.i.d. across agents, the ex-ante expected lottery for agent 1 under $f_{31}$ must be the same as that of agent 3. Therefore, both agents 1 and 3 each gets good $c$ with ex-ante expected probability of $p_c(\delta) > 1/2$. However, the ex-ante expected assignment is a convex combination of the ex-post (random) assignments, so the sum of the probabilities for each of the goods must be exactly 1, a contradiction. □

## Appendix B. Independent types and interdependent values example

Let $n = 2$, $O = \{x, y\}$, and $T = \{G, R\}$. Types are independent with the common prior being $\mu(R,R) = \mu(R,G) = \mu(G,R) = \mu(G,G) = 1/4$. Both agents have the same utility function given by the following matrices, where rows correspond to the agent's own types, columns to the other agent types, and each matrix to a different alternative:

$x$:

| | $G$ | $R$ |
|---|---|---|
| $G$ | 0 | 1 |
| $R$ | 0 | 0 |

$y$:

| | $G$ | $R$ |
|---|---|---|
| $G$ | 1 | 0 |
| $R$ | 1 | 1 |

In words, each agent prefers alternative $y$ over $x$ in all type profiles, except when his type is $G$ and the other agent type is $R$, in which case he prefers $x$ over $y$.

Consider the SCF $f$ given by the following matrix:

| | $G$ | $R$ |
|---|---|---|
| $G$ | $x$ | $y$ |
| $R$ | $x$ | $y$ |

This is a dictatorial SCF with agent 2 being the dictator. It is clearly incentive compatible for agent 1. It is also incentive compatible for agent 2, since he believes that agent 1 is equally likely to be of type $G$ or $R$, so if his own type is $G$ then he is indifferent between $x$ and $y$, and if his own type is $R$ then he prefers $y$.

We claim that $f$ cannot be symmetrically implemented in BNE, and we use the condition in Theorem 1 to prove this. Consider agent 1 of type $G$. His expected utility from truthfully reporting his type is 0. Thus, for symmetric implementation we need to find probabilities (for outcome $x$) $f_{12}(G,G)$, $f_{12}(G,R) = f_{12}(R,G)$, and $f_{12}(R,R)$ such that

$$0 \geq \frac{1}{2}\left[f_{12}(G,G) * 0 + (1 - f_{12}(G,G)) * 1\right] + \frac{1}{2}\left[f_{12}(G,R) * 1 + (1 - f_{12}(G,R)) * 0\right],$$

and

$$0 \geq \frac{1}{2}\left[f_{12}(R,G) * 0 + (1 - f_{12}(R,G)) * 1\right] + \frac{1}{2}\left[f_{12}(R,R) * 0 + (1 - f_{12}(R,R)) * 1\right].$$

However, the first inequality implies that $f_{12}(G,R) = 0$ and the second implies that $f_{12}(R,G) = 1$, a contradiction.