Theoretical Economics 9 (2014), 339–360 1555-7561/20140339

Robust mechanism design and dominant strategy voting rules

Tilman Börgers
Department of Economics, University of Michigan

Doug Smith
Federal Trade Commission

We develop an analysis of voting rules that is robust in the sense that we do not make any assumption regarding voters' knowledge about each other. In dominant strategy voting rules, voters' behavior can be predicted *uniquely* without making any such assumption. However, on full domains, the only dominant strategy voting rules are random dictatorships. We show that the designer of a voting rule can achieve Pareto improvements over random dictatorship by choosing rules in which voters' behavior can depend on their beliefs. The Pareto improvement is achieved for all possible beliefs. The mechanism that we use to demonstrate this result is simple and intuitive, and the Pareto improvement result extends to all equilibria of the mechanism that satisfy a mild refinement. We also show that the result only holds for voters' interim expected utilities, not for their ex post expected utilities.

Keywords. Robust mechanism design, Gibbard–Satterthwaite theorem.

JEL classification. D71, D82.

1. Introduction

In this paper, we consider the design of voting rules from the perspective of the theory of robust mechanism design. Our starting point is the classic result due to Gibbard (1973) and Satterthwaite (1975) according to which the only dominant strategy voting rules for three or more alternatives are dictatorial voting rules. Gibbard and Satterthwaite assumed the number of alternatives to be finite. Preferences were modeled as complete and transitive orders of the set of alternatives. For every voter, the range of relevant preferences was taken to be the set of *all* possible preferences over the alternatives. Gibbard and Satterthwaite then asked whether it is possible to construct a game form[1] that determines which alternative is selected as a function of the strategies chosen by the voters, such that each voter has a dominant strategy regardless of what that voter's preferences

Tilman Börgers: tborgers@umich.edu
Doug Smith: dsmith2@ftc.gov

This paper is a chapter of Doug Smith's dissertation at the University of Michigan. We thank Stephan Lauermann, Arunava Sen, a co-editor, and two referees for very helpful comments. The views expressed in this article are those of the authors and do not necessarily reflect those of the Federal Trade Commission.

[1] We use the terms "game form" and "mechanism" synonymously.

DOI: 10.3982/TE1111

are. A dominant strategy was defined to be a strategy that is a best reply to each of the other voters' strategy combinations. Gibbard and Satterthwaite also required that every alternative be the outcome under at least one strategy profile. They then showed that the only game forms that satisfy this requirement, and that offer each voter, for every preference of that voter, a dominant strategy are game forms that leave the choice of the outcome to just one individual, the dictator.[2]

One motivation for the interest in dominant strategy mechanisms is that dominant strategies predict rational voters' behavior without relying on any assumption about the voters' beliefs about each others' preferences or behavior. If a voter does not have a dominant strategy, then that voter's optimal choice depends on his beliefs about other voters' behavior, which in turn may be derived from beliefs about other voters' preferences. It seems at first sight attractive to bypass such beliefs and to construct a game form for which a prediction can be made that is independent of beliefs.

On closer inspection, this argument can be seen to consist of two parts:

A. *The design of a good game form for voting should not be based on specific assumptions about voters' beliefs about each other.*

B. *A good game form for voting should allow us to predict rational voters' choices uniquely from their preferences, without making specific assumptions about these voters' beliefs about each other.*[3]

Both parts of the argument have their own appeal. Voting schemes are often constructed long before the particular contexts in which they will be used are known. It seems wise not to make any special assumptions about agents' knowledge about each other, motivating part A of the argument. Part B can perhaps be motivated by the idea that game forms in which voters' behavior can be uniquely predicted independently of their beliefs confront voters with simpler strategic problems than game forms in which voters' rational behavior is belief-dependent.

As the Gibbard–Satterthwaite theorem shows, A and B together impose strong restrictions on a voting scheme. In this paper we maintain A, but drop B. In other words, we examine game forms for voting without making assumptions about voters' beliefs about each other, but we do not restrict attention to game forms for which voters' equilibrium strategies are independent of those beliefs. Our work is thus in a sense complementary to the work surveyed by Barberà (2010) that insists on dominant strategies, requirement B, but seeks to obtain more positive results than Gibbard and Satterthwaite by restricting the domain of preferences that is considered.

To be able to use the notion of *Bayesian equilibrium* in our formal analysis, we need to introduce a framework that is slightly different from Gibbard and Satterthwaite's

[2]The literature that builds on Gibbard and Satterthwaite's seminal work is voluminous. For a recent survey, see Barberà (2010).

[3]Blin and Satterthwaite (1977) emphasize the interpretation of the Gibbard–Satterthwaite theorem as a result about voting procedures in which each voter's choice depends only on his/her preferences and not on his/her beliefs about others' preferences.

framework. We model voters' attitudes toward risk, assuming that they maximize expected utility. A version of Gibbard and Satterthwaite's theorem for expected utility maximizers was shown by Hylland (1980), who assumed that voters have von Neumann–Morgenstern utilities and that lotteries are allowed as outcomes. He characterized game forms that offer each agent a dominant strategy for every utility function, that pick an alternative with probability 1 if it is unanimously preferred by all agents, and that pick an alternative with probability 0 if it is unanimously ranked lowest by all agents. He showed that, when there are at least three alternatives, the only such game forms are random dictatorships.[4] In random dictatorships, each voter i gets to be dictator with a probability p_i that is independent of all preferences. If voter i is dictator, then the outcome that voter i ranks highest is chosen.

The two main results of this paper address whether there are game forms such that, for all type spaces, there is at least one Bayesian equilibrium of the game form that yields all voters' types the same expected utility and, for some voters' types in some type spaces, strictly higher expected utility than random dictatorship. Obviously, the answer to this question can be positive only when each voter's probability of being dictator is strictly less than 1. In our first main result, we show that in this case the answer to our question is indeed positive, provided that we consider interim expected utility, that is, each voter's expected utility is calculated when that voter's type is known, but the other voters' types are not yet known.[5] If an ex post perspective is adopted instead, that is, if voters' expected utility is considered conditional on the vector of *all* voters' types, then no voting game form Pareto improves on random dictatorship. This is our second main result.

We show the first main result using a simple game form that allows voters to avoid random dictatorship and implement a compromise whenever all voters agree that the compromise is preferable to random dictatorship. It will be easy to see that our first main result can be extended and that we can show that not just one, but *all* Bayesian equilibria of the game form that we are proposing, if they satisfy a mild refinement, yield for all voters at least as high expected utility as random dictatorship.[6]

The compromise option may not turn out to be a Pareto improvement ex post, as agents may compromise because they think it likely that the compromise will improve on random dictatorship, but ex post discover to have a type vector that appeared unlikely, and for which the compromise is not a Pareto improvement. The second main result shows that *any* game form other than random dictatorship will sometimes make some type worse off in comparison to random dictatorship.

[4]This result is Theorem 1* in Hylland (1980). It is also Theorem 1 in Dutta et al. (2007) (see also Dutta et al. 2008), where an alternative proof is provided. Another proof is given in Nandeibam (2013).

[5]The notions of interim and ex post efficiency are due to Holmström and Myerson (1983).

[6]Our first, positive result is thus in the spirit of the literature on *full implementation*, which considers all equilibria of a game form, whereas our second, negative result is in the spirit of the literature on *mechanism design*, which considers only some equilibrium of a given game form. Both results are stronger than they would be if the respective other approach were used. For the distinction between implementation and mechanism design, see, for example, Jackson (2001).

Important limitations of our work are that we only consider *finite* type spaces and that we require Bayesian equilibria to be *consistent*. We define "consistency" in this paper to mean that equilibrium actions, although they may depend on agents' beliefs, do not depend on the details of the formal representation of those beliefs. That we only consider finite type spaces makes the first, positive result weaker, but it makes the second, negative result stronger than it would be otherwise. The use of an equilibrium refinement makes the first result stronger, but it makes the second result weaker than it would be if we allowed all Bayesian equilibria.

We do not know whether our first, positive result would remain true if we considered large type spaces such as the universal type space constructed by Mertens and Zamir (1985) or a similar universal type space instead of the set of all finite type spaces. Our second, negative result would not remain true if we allowed all Bayesian equilibria. Our consistency requirement is a very weak requirement, however, and we believe it to be persuasive. We discuss the details of the points that we have just touched on later in the paper.

Section 2 discusses related literature. Section 3 explains the model and the definitions used in this paper. In Section 4, we adapt Hylland's theorem on random dictatorship to our setting. In Section 5, we explain how we relax the requirement that voters' choices, for given preferences, are the same in all type spaces. Sections 6 and 7 contain our two main results. Section 8 concludes.

2. Related literature

Our approach builds on Bergemann and Morris' (2005) seminal work on robust mechanism design.[7] They consider, as we do, Bayesian equilibria of mechanisms on *all* type spaces. Unlike us, they do not rule out infinite type spaces and they do not require equilibria to be consistent. Bergemann and Morris study sufficient conditions under which the Bayesian implementability of a social choice correspondence on all type spaces implies dominant strategy implementability (or, more generally, implementability in *ex post equilibria*). The main sufficient condition that they find is that the economic environment is *separable*. The two prime examples of separable environments are environments in which the social choice correspondence is singleton valued, and environments in which each agent's utility depends quasilinearly on a common component of the outcome and the individual agent's monetary transfer.

When strategies in a Bayesian equilibrium are belief-independent in the sense of requirement B in the Introduction and all type spaces are considered, then the implemented social choice correspondence is obviously singleton valued. Moreover, implementation of a singleton social choice correspondence on all type spaces implies that truthful revelation of preferences is a dominant strategy, not only if infinite type spaces are included, as in Bergemann and Morris, but also if only finite type spaces are considered and also if attention is restricted to consistent Bayesian equilibria. We show this

[7] The literature on robust mechanism design and implementation was recently surveyed by Bergemann and Morris (2012).

simple observation in the proof of the preliminary result, Proposition 1 below, and use it in that proof to apply Hylland's theorem to social choice correspondences implemented in belief-independent Bayesian equilibria for all finite type spaces.

Bergemann and Morris (2005, Section 6.3) point out that in nonseparable environments, such as the environment without transferrable payoffs that we consider, dominant strategy implementability may be a stronger requirement than Bayesian implementability on all type spaces.[8] Our paper shows that this observation remains true when only consistent Bayesian equilibria are allowed and only finite type spaces are considered.

Unlike our paper, Bergemann and Morris' work, by restricting attention to social choice correspondences, considers only ex post, but not interim, normative criteria. Moreover, Bergemann and Morris do not consider how a mechanism designer compares different mechanisms if none of them implements the social choice correspondence that describes the designer's most preferred outcomes. Such comparisons are the focus of our work.

The approach to comparing different mechanisms that we take in this paper is based on Smith (2010), who studies the design of a mechanism for public goods. Smith considers the performance of different mechanisms in a Bayesian equilibrium on all type spaces. He focuses on an ex post perspective and demonstrates that a mechanism designer can improve efficiency using a more flexible mechanism than a dominant strategy mechanism. In our setting, by contrast, we find that no mechanism can improve on dominant strategy mechanisms ex post, but that such an improvement is possible from an interim perspective. Smith finds improvement possibilities for *any* dominant strategy mechanism, whereas in our setting improvements are only possible for those dominant strategy mechanisms where the probability that any specific agent's action solely determines the outcome is less than 1.

Chung and Ely (2007) describe an auctioneer of a single object who designs an auction to maximize expected revenues. The auctioneer considers equilibria of different auction mechanisms on the universal type space and evaluates different mechanisms using a maximin criterion: taking the distribution of the agents' valuations, but not the agents' beliefs, as given, for each mechanism, the auctioneer determines the probability distribution on the universal type space for which that mechanism yields the lowest expected revenue. The auctioneer then chooses a mechanism that maximizes the lowest expected revenue. Aside from the obvious differences in setting, the main conceptual difference from our work is that our mechanism designer has only a partial order of mechanisms, whereas Chung and Ely's mechanism designer has a complete order. Our order is based on comparing mechanisms on every type space, and ranking one mechanism above another if it performs according to the designer's objectives at least as well on all type spaces and better on some. For this order, we find, unlike Chung and Ely, that in our setting there are mechanisms that perform better than dominant strategy mechanisms.

[8]The discussion paper version (Bergemann and Morris 2003) of Bergemann and Morris (2005) also includes a general characterization of Bayesian implementability on all type spaces; however, we do not make use of this characterization.

Whereas the papers cited so far are concerned with mechanism design, in the sense that for any given mechanism and type space only one Bayesian equilibrium is considered, there is also a literature on robust implementation, in which, for any given mechanism and type space, all Bayesian equilibria are taken into account. Bergemann and Morris (2011) provide conditions for a social choice function to be implementable on every type space.

A recent paper by Yamashita (2012) is related to the idea of robust implementation. Yamashita considers a bilateral trade setting and evaluates mechanisms on the basis of the lowest expected welfare among all outcomes that can result if agents use strategies that are not weakly dominated. Expected welfare is calculated on the basis of the mechanism designer's subjective prior over agents' types. Yamashita's work is similar to work on implementation because he considers all outcomes, not just some outcomes, that can result under a solution concept. A predecessor to Yamashita (2012) is Börgers (1991), who considers, in the Gibbard–Satterthwaite framework, the existence of mechanisms for which the outcomes that result if all players choose a strategy that is not weakly dominated are Pareto efficient and (in a sense defined in that paper) less one-sided than the outcomes of dictatorship. Börgers shows the existence of such mechanisms. Börgers uses a framework in which agents' preferences are modeled using ordinal preferences rather than von Neumann Morgenstern utilities.[9]

In Börgers and Smith (2012), we further develop the approach of implementation in non-weakly-dominated strategies. Among the applications that we consider are voting mechanisms. We show the possibility of an ex post improvement over random dictatorship if one evaluates outcomes by taking the expected value of a Rawlsian welfare function.

Bayesian mechanism design approaches to voting are surprisingly rare in the literature. For the case of independent types, Azrieli and Kim (forthcoming) recently considered interim and ex ante efficiency in a setting with two alternatives and independent types. Schmitz and Tröger (2012) consider the same issue and allow correlated types. Börgers and Postl (2009) study ex ante welfare maximization in a setting with three alternatives. The type space in their paper is very small, with the ordinal ranking of alternatives being common knowledge and only the cardinal utility functions being private information.

The game form that we use to prove our first main result—that random dictatorship can be improved on from an interim perspective—is almost identical to the *full consensus or random ballot fall-back* game form that Heitzig and Simmons (2012) introduced. While their motivation, like ours, is to consider voting systems that are more flexible than dictatorial voting systems and that allow for compromises, the focus of their formal analysis is on complete information, correlated equilibria that are in some sense coalition-proof. In our paper, the focus is on analyzing Bayesian equilibria in arbitrary type spaces.

[9]Theorem III in Zeckhauser (1973) is a related but negative result for the case of von Neumann–Morgenstern utilities.

3. The voting problem

There are n agents: $i \in I = \{1, 2, \dots, n\}$. The agents have to choose one alternative from a finite set A of alternatives that has at least three elements. The set of all probability distributions over A is $\Delta(A)$, where, for $\delta \in \Delta(A)$ and $a \in A$, we denote by $\delta(a) \in [0, 1]$ the probability that δ assigns to a. The agents are commonly known to be expected utility maximizers. We denote agent i's von Neumann–Morgenstern utility function by $u_i : A \to \mathbb{R}$. We assume that $a \neq b \Rightarrow u_i(a) \neq u_i(b)$, that is, there are no indifferences. We define the expected utility for probability distributions $\delta \in \Delta(A)$ by $u_i(\delta) = \sum_{a \in A} u_i(a)\delta(a)$.

A mechanism designer has a (possibly incomplete) ranking of the alternatives in A that may depend on the agents' utility functions. We shall be more specific about the designer's objectives later. The mechanism designer does not know the agents' utility functions; neither does she know what the agents believe about each other. To implement an outcome that potentially depends on the agents' utility functions, the mechanism designer asks the agents to play a game form.

Definition 1. A *game form* $G = (S, x)$ has two components:

(i) A set $S \equiv \prod_{i \in I} S_i$, where for every $i \in I$, the set S_i is nonempty and finite.

(ii) A function $x : S \to \Delta(A)$.

The set S_i is the set of (pure) strategies available to agent i in the game form G. We focus on finite sets of pure strategies, while allowing mixed strategies, to ease exposition. Our results also hold when the sets S_i of pure strategies are allowed to be infinite. The function x assigns to every combination of pure strategies s the potentially stochastic outcome $x(s)$ that is implemented when agents choose that combination of pure strategies. We write $x(s, a)$ for the probability that $x(s)$ assigns to alternative a.

Once the mechanism designer has announced a game form, the agents choose their strategies simultaneously and independently. Because the agents do not necessarily know each others' utility functions or beliefs, this game may be a game of incomplete information. A hypothesis about the agents' utility functions and their beliefs about each other can be described by a type space.[10]

Definition 2. A *type space* $\mathcal{T} = (T, \pi, u)$ has the following components:

(i) A set $T \equiv \prod_{i \in I} T_i$, where, for every $i \in I$, the set T_i is nonempty and finite.

(ii) An array $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ of functions $\pi_i : T_i \to \Delta(T_{-i})$, where $\Delta(T_{-i})$ is the set of all probability distributions over $T_{-i} \equiv \prod_{j \neq i} T_j$.

(iii) An array $u = (u_1, u_2, \dots, u_n)$ of functions $u_i : T_i \times A \to \mathbb{R}$ such that $a \neq b \Rightarrow u_i(t_i, a) \neq u_i(t_i, b)$ for all $t_i \in T_i$.

[10] The following definition only refers to finite type spaces. To simplify the terminology, we omit the adjective "finite," but we discuss the role that finiteness plays in our analysis after explaining the definition and also later in the paper.

The set T_i is the set of types of agent i. Agent i privately observes his type $t_i \in T_i$. The function π_i describes, for every type $t_i \in T_i$, the beliefs that agent i has about the other agents' types when agent i himself is of type t_i. We write $\pi_i(t_i, t_{-i})$ for the probability that type t_i assigns to the other players types being t_{-i}. Beliefs are subjective. There may or may not be a common prior for a particular type space. Different agents' beliefs may be incompatible with each other in the sense that one agent may attach positive probability to an event to which another agent attaches probability zero. The function $u_i(t_i)$ describes player i's utility when i is of type t_i. We write $u_i(t_i, a)$ for the utility that $u_i(t_i)$ assigns to alternative a. The utility functions $u_i(t_i)$ satisfy the assumption introduced earlier that there are no indifferences.[11] We allow redundant types, that is, multiple types with identical utility functions and identical hierarchies of beliefs over all players' utility functions. The possible importance of redundant types for the analysis of Bayesian equilibria has been emphasized by Ely and Pęski (2006, Section 1.2).

We assume that the mechanism designer has no knowledge of the agents' utility functions or their beliefs. Therefore, the mechanism designer regards all type spaces as possible descriptions of the environment. We denote the set of all type spaces by Υ.[12] Alternatively, one may think of Υ as just one large type space.

Note that we have assumed the sets T_i to be finite. Therefore, type spaces such as the universal type space constructed by Mertens and Zamir (1985) or by Sadzik (2011)[13] are not contained in Υ; neither is Υ in some sense equivalent to a universal type space. To see the last point, note that every type in any type space in Υ believes that it is common knowledge among agents that the cardinality of the support of all agents' beliefs at any level of their belief hierarchy is finite, whereas this is not the case for every type in the universal type space. Our construction thus rules out some hierarchies of beliefs that are allowed by either of the two universal type spaces mentioned above. We do not know whether Proposition 2 below would remain true if we considered an appropriate universal type space. We shall explain the difficulty in extending Proposition 2 to universal type spaces after the proof of Proposition 2. All other results in this paper would remain unchanged if we considered either of the two universal type spaces mentioned above.

The mechanism designer proposes to agents how they might play the game. For the agents to accept the mechanism designer's proposal, she must propose a *Bayesian equilibrium*. Because the mechanism designer does not know the true type space, she has to propose a *Bayesian equilibrium for every type space*.

Definition 3. A *Bayesian equilibrium of game form G for every type space* is an array $\sigma^* = (\sigma_1^*, \sigma_2^*, \ldots, \sigma_n^*)$ such that for every $i \in I$,

(i) σ_i^* is a family of functions $(\sigma_i^*(\mathcal{T}))_{\mathcal{T}\in\Upsilon}$, where for every $\mathcal{T} \in \Upsilon$, the function $\sigma_i^*(\mathcal{T})$ maps the type space T_i that corresponds to $\mathcal{T}$ into $\Delta(S_i)$, the set of all probability distributions on S_i,

[11] Observe that we suppress in the notation the dependence of π_i and u_i on the type space $\mathcal{T}$. No confusion should arise from this simplification of our notation.

[12] More precisely, Υ is the set of all *finite* type spaces.

[13] Sadzik (2011) constructs a universal type space that is appropriate for the study of Bayesian Nash equilibria if one wants to allow redundant types.

and, writing $\sigma_i^*(\mathcal{T}, t_i)$ for the mixed strategy assigned to t_i and writing $\sigma_i^*(\mathcal{T}, t_i, s_i)$ for the probability that this mixed strategy assigns to $s_i \in S_i$, we have for every $\mathcal{T} \in \Upsilon$, $i \in I$, and $t_i \in T_i$ (where T_i corresponds to $\mathcal{T}$),

(ii) $\sigma_i^*(\mathcal{T}, t_i)$ maximizes the expected utility of type t_i among all mixed strategies in $\Delta(S_i)$, where expected utility for any mixed strategy $\sigma_i \in \Delta(S_i)$ is

$$\sum_{t_{-i} \in T_{-i}} \pi_i(t_i, t_{-i}) \sum_{s \in S} u_i(t_i, x(s)) \cdot \sigma_i(s_i) \cdot \prod_{j \neq i} \sigma_j^*(\mathcal{T}, t_j, s_j).$$

The mechanism designer evaluates different mechanisms and their equilibria using the Pareto criterion. When evaluating the agents' utility for a realized type combination t, the mechanism designer might only consider the outcomes that result from the mixed strategies prescribed for these types. Alternatively, the mechanism designer might consider the expected utilities of these types, based on the types' own subjective beliefs. In other words, the mechanism designer may adopt an ex post or an interim perspective when evaluating agents' utilities. The interim perspective respects agents' own perception of their environment. The ex post perspective has a paternalistic flavor. On the other hand, for example, when agents' beliefs are incompatible with each other, the mechanism designer may be justified in discarding agents' beliefs on the basis that at least some of them have to be wrong, as agents themselves will discover at some point. Thus neither the interim nor the ex post perspective seems clearly preferable. We pursue both perspectives in this paper.

DEFINITION 4. The game form G and the Bayesian equilibrium for all type spaces σ^* *interim Pareto dominate* the game form $\widetilde{G}$ and the Bayesian equilibrium for all type spaces $\tilde{\sigma}^*$ if for all $\mathcal{T} \in \Upsilon$, $i \in I$, and $t_i \in T_i$,

$$\sum_{t_{-i} \in T_{-i}} \pi_i(t_i, t_{-i}) \sum_{s \in S} u_i(t_i, x(s)) \cdot \prod_{j \in I} \sigma_j^*(\mathcal{T}, t_j, s_j) \geq \sum_{t_{-i} \in T_{-i}} \pi_i(t_i, t_{-i}) \sum_{s \in \widetilde{S}} u_i(t_i, \tilde{x}(s)) \cdot \prod_{j \in I} \tilde{\sigma}_j^*(\mathcal{T}, t_j, s_j),$$

with strict inequality for at least one $\mathcal{T} \in \Upsilon$, $i \in I$, and $t_i \in T_i$.

DEFINITION 5. The game form G and the Bayesian equilibrium for all type spaces σ^* *ex post Pareto dominate* the game form $\widetilde{G}$ and the Bayesian equilibrium for all type spaces $\tilde{\sigma}^*$ if for all $\mathcal{T} \in \Upsilon$, $i \in I$, and $t \in T$,

$$\sum_{s \in S} u_i(t_i, x(s)) \cdot \prod_{j \in I} \sigma_j^*(\mathcal{T}, t_j, s_j) \geq \sum_{s \in \widetilde{S}} u_i(t_i, \tilde{x}(s)) \cdot \prod_{j \in I} \tilde{\sigma}_j^*(\mathcal{T}, t_j, s_j),$$

with strict inequality for at least one $\mathcal{T} \in \Upsilon$, $i \in I$, and $t \in T$.

Our main interest in this paper is to explore how the mechanism designer's ability to achieve her objective depends on additional conditions that Bayesian equilibria

of the mechanism designer's proposed game form have to satisfy. In the next section, we consider the very restrictive requirement of belief independence. In the subsequent sections, we relax this requirement.

4. Belief-independent equilibria: Hylland's theorem

We begin by exploring the consequences of a restrictive requirement for the Bayesian equilibrium that the mechanism designer proposes. This requirement is implicit in the work on dominant strategy mechanism design. It is that equilibria be *belief-independent*. Using the notion of belief-independent equilibria, we can restate Hylland's version of the Gibbard–Satterthwaite theorem in our setting.

Definition 6. A Bayesian equilibrium for every type space, σ^*, of a game form G is *belief-independent* if for all $i \in I$, $\mathcal{T}, \widetilde{\mathcal{T}} \in \Upsilon$, $t_i \in T_i$, and $\tilde{t}_i \in \widetilde{T}_i$ such that $u_i(t_i) = \tilde{u}_i(\tilde{t}_i)$, we have

$$\sigma_i^*(\mathcal{T}, t_i) = \sigma_i^*(\widetilde{\mathcal{T}}, \tilde{t}_i),$$

where T_i, u_i correspond to $\mathcal{T}$ and $\widetilde{T}_i, \tilde{u}_i$ correspond to $\widetilde{\mathcal{T}}$.

The reformulation of Hylland's theorem presented below says that all game forms and belief-independent equilibria of these game forms that satisfy two unanimity requirements are random dictatorships. To define the two unanimity requirements and random dictatorship, we need some notation. If u is a utility function, we denote by $b(u)$ the element of A that maximizes u and denote by $w(u)$ the element of A that minimizes u.[14]

Definition 7. A game form G and a Bayesian equilibrium of G for every type space, σ^*, satisfy:

(i) *Positive unanimity* if for every $\mathcal{T} \in \Upsilon$, $t \in T$, and $a \in A$ such that $b(u_i(t_i)) = a$ for all $i \in I$, we have

$$\sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s, a) = 1.$$

(ii) *Negative unanimity* if for every $\mathcal{T} \in \Upsilon$, $t \in T$, and $a \in A$ such that $w(u_i(t_i)) = a$ for all $i \in I$, we have

$$\sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s, a) = 0.$$

Positive and negative unanimity are implied by, but weaker than, ex post Pareto efficiency. Next, we provide the formal definition of random dictatorship that we need for our reformulation of Hylland's theorem.

[14] Recall that we have assumed that there are no indifferences. Therefore, there is a unique element of A that maximizes u and a unique element of A that minimizes u.

DEFINITION 8. A game form G and a Bayesian equilibrium of G for every type space, σ^*, are a *random dictatorship* if there is some $p \in [0,1]^n$ such that for every $\mathcal{T} \in \Upsilon$, $t \in T$, and $a \in A$,

$$\sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s,a) = \sum_{\{i \in I : b(u_i(t_i)) = a\}} p_i.$$

The following proposition is implied by Hylland's theorem.[15]

PROPOSITION 1. *A game form G and a belief-independent Bayesian equilibrium of G for every type space, σ^*, satisfy positive and negative unanimity if and only if they are a random dictatorship.*

PROOF. The "if" part is obvious. To prove the "only if" part, we derive from G and σ^* a "cardinal decision scheme" in the sense of Definition 1 in Dutta et al. (2007) and show that this cardinal decision scheme has the properties listed in Theorem 1 in Dutta et al. (2007) and the correction in Dutta et al. (2008). It then follows from Theorem 1 in Dutta et al. (2007) that the cardinal decision scheme is a random dictatorship. This then implies the "only if" part of our Proposition 1.

Denote by $\mathcal{U}$ the set of all utility functions that have the property of no indifferences (see Definition 2). A cardinal decision scheme is a mapping $\phi : \mathcal{U}^n \to \Delta(A)$. We can derive from G and σ^* a cardinal decision scheme by setting, for any $(u_1, u_2, \ldots, u_n) \in \mathcal{U}^n$ and $a \in A$, the probability $\phi(u_1, u_2, \ldots, u_n, a)$ that $\phi(u_1, u_2, \ldots, u_n)$ assigns to a as

$$\phi(u_1, u_2, \ldots, u_n, a) = \sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s,a),$$

where we can pick any $\mathcal{T} \in \Upsilon$ and any $t \in T$ such that $u_i(t_i) = u_i$ for all $i \in I$. By belief independence, it does not matter which such $\mathcal{T}$ and $t \in T$ we choose. Then ϕ is a cardinal decision scheme as defined in Definition 1 of Dutta et al. (2007).

We can complete the proof by showing that ϕ has the two properties listed in Theorem 1 of Dutta et al. (2007) and the additional property listed in the correction (Dutta et al. 2008). The first property is unanimity: If $b(u_i) = a$ for all $i \in I$, then $\phi(u_1, u_2, \ldots, u_n, a) = 1$. This is implied by the assumption that G and σ^* satisfy positive unanimity.

The second property is strategy proofness: If $(u_1, u_2, \ldots, u_n) \in \mathcal{U}^n$ and $u_i' \in \mathcal{U}$, then $u_i(\phi(u_i, u_{-i})) \geq u_i(\phi(u_i', u_{-i}))$, where u_{-i} is the array $(u_1, u_2, \ldots, u_n)$ that leaves out u_i. To prove this, we pick $\mathcal{T} \in \Upsilon$, $t_i, t_i' \in T_i$, and $t_{-i} \in \prod_{j \neq i} T_j$ such that $u_i(t_i) = u_i$, $u_i(t_i') = u_i'$, and $u_j(t_j) = u_j$ for all $j \neq i$. Moreover, $\pi_i(t_i)$ and $\pi_i(t_i')$ place probability 1 on t_{-i}. Then the fact that σ^* is a Bayesian equilibrium of G for the type space $\mathcal{T}$ implies

$$\sum_{s \in S} u_i(t_i, x(s)) \cdot \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot \prod_{j \neq i} \sigma_j^*(\mathcal{T}, t_j, s_j) \geq \sum_{s \in S} u_i(t_i, x(s)) \cdot \sigma_i^*(\mathcal{T}, t_i', s_i) \cdot \prod_{j \neq i} \sigma_j^*(\mathcal{T}, t_j, s_j).$$

[15] Theorem 1* in Hylland (1980). We use here the version of Hylland's theorem that is Theorem 1 in Dutta et al. (2007) with the correction in Dutta et al. (2008).

By the definition of ϕ, this is equivalent to $u_i(\phi(u_i, u_{-i}) \geq u_i(\phi(u_i', u_{-i}))$, that is, strategy proofness.

The third property, introduced in the correction (Dutta et al. 2008), is a property labelled (*) in Dutta et al. (2008): If $w(u_i) = a$ for all $i \in I$, then $\phi(u_1, u_2, \ldots, u_n, a) = 0$. This is implied by the assumption that G and σ^* satisfy negative unanimity. □

From now on, when we refer to random dictatorship, we mean a specific game form G and a specific equilibrium σ^* of G for every type space.

Definition 9. For any vector $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$, the following game form G and equilibrium σ^* of G for every type space will be referred to as *p-random dictatorship*:

(i) $S_i = A$ for all $i \in I$

(ii) $x(s,a) = \sum_{\{i\in I: s_i = a\}} p_i$ for all $s \in S$ and $a \in A$

(iii) $\sigma_i^*(\mathcal{T}, t_i, b(u_i(t_i))) = 1$ for all $i \in I$, $\mathcal{T} \in \Upsilon$, and $t_i \in T_i$.

It is immediate that σ^* is a belief-independent Bayesian equilibrium of G for every type space, and that G and this equilibrium are a random dictatorship. There are other game forms and equilibria that are random dictatorships, but it is without loss of generality to only consider the ones described in Definition 9.

5. Consistent equilibria

Our main interest in this paper is to consider the implications of relaxing the requirement of belief independence for the Bayesian equilibria of the game form that the mechanism designer chooses. We do not, however, completely dispense with any link between players' strategies in different type spaces. The Bayesian equilibria that we shall investigate need to satisfy a *consistency* requirement. This requirement is implied by, but does not imply, belief independence.

Definition 10. A Bayesian equilibrium of game form G for every type space, σ^*, is *consistent* if for all type spaces $\mathcal{T}, \widetilde{\mathcal{T}} \in \Upsilon$, if the following statements hold:

(i) for every $i \in I$, $T_i \subseteq \widetilde{T}_i$ (where T_i corresponds to $\mathcal{T}$ and $\widetilde{T}_i$ corresponds to $\widetilde{\mathcal{T}}$),

(ii) for every $i \in I$ and every $t_i \in T_i$, $\tilde{u}_i(t_i) = u_i(t_i)$ and $\tilde{\pi}_i(t_i) = \pi_i(t_i)$ (where u_i, π_i correspond to $\mathcal{T}$ and $\tilde{u}_i$, $\tilde{\pi}_i$ correspond to $\widetilde{\mathcal{T}}$),

then

(iii) we have for every $i \in I$ and every $t_i \in T_i$, $\sigma_i^*(\widetilde{\mathcal{T}}, t_i) = \sigma_i^*(\mathcal{T}, t_i)$.

While Proposition 2 below remains true even if one considers all Bayesian equilibria, not just consistent equilibria, Proposition 3 does not. Proposition 3 does remain true, however, if we include a universal type space among the type spaces that we consider.

We discuss these points in Section 7. In any case, we regard the consistency refinement as eminently plausible. Observe that the type t_i referred to in item (iii) of Definition 10 has the same utility function and hierarchy of beliefs over other players' utility functions and types in type space $\widetilde{\mathcal{T}}$ as in type space $\mathcal{T}$. In technical language, type space $\mathcal{T}$ is a "belief-closed subspace" of type space $\widetilde{\mathcal{T}}$. Enlarging the type space from $\mathcal{T}$ to $\widetilde{\mathcal{T}}$ does not reflect any change in the beliefs of types in type space $\mathcal{T}$, but instead reflects that the modeler or the mechanism designer considers more specifications of agents' beliefs than were included in type space $\mathcal{T}$. Therefore, agents' strategy choices for types in $\mathcal{T}$ should not change when the type space is enlarged from $\mathcal{T}$ to $\widetilde{\mathcal{T}}$. This is the content of the consistency refinement.

If one interprets a player's type in a type space as a convenient representation of that player's hierarchy of beliefs, then consistency is a very cautious formalization of the requirement that the particular representation of this hierarchy of beliefs should not matter as long as the hierarchy itself is unchanged. This requirement is an example of the invariance requirements studied in Yildiz (2011). These invariance requirements place the restriction on the selection of equilibria of the same game for different type spaces that types with identical information play the same equilibrium action. As Yildiz emphasizes, there are different reasonable interpretations of the phrase "identical information" and corresponding different invariance requirements. In Definition 10, we interpret identical information to mean identical hierarchies of beliefs about players' utility function and types.[16]

Note that we allow equilibrium actions to depend on the labels of players' types, which makes consistency a particularly weak requirement. In particular, consistency does not imply that equilibrium actions are the same for redundant types, that is, equilibrium actions need not only depend on a player's utility function and the player's hierarchy of beliefs about players' utility functions, that is, the player's hierarchy of beliefs in Mertens and Zamir's universal type space. However, when a Bayesian equilibrium of a game on Mertens and Zamir's universal type space exists, one can construct a corresponding consistent equilibrium for all finite type spaces by appealing to the "equilibrium pull-back property" of Friedenberg and Meier (2010, Proposition 4.1).

6. A game form that interim Pareto dominates random dictatorship

The first main result of this paper examines interim Pareto dominance, while the second main result concerns ex post Pareto dominance. The first result says that for every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$ and $p_i < 1$ for all $i \in I$, there is a game form and a consistent equilibrium of this game form for every type space that interim Pareto dominates p-random dictatorship.[17] We refer to the dominating game form as *p-random dictatorship with compromise*.

[16] One checks easily that our invariance notion has the property that Yildiz requires invariance notions to have.

[17] If $p_i = 1$ for some $i \in I$, that is, when dictatorship is deterministic, not random, then obviously no game form can interim Pareto dominate dictatorship.

DEFINITION 11. For every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$, the following game form is called a *p-random dictatorship with compromise*.

(i) For every $i \in I$,

$$S_i = \mathcal{A} \times \mathcal{R},$$

where $\mathcal{A}$ is the set of all nonempty subsets of A and $\mathcal{R}$ is the set of all complete strict ordinal rankings of A. We write $s_i = (A_i, R_i) \in S_i$ for a strategy for agent i.

(ii) If $\bigcap_{i\in I} A_i = \varnothing$, then for all $a \in A$,

$$x(s,a) = \sum_{\{i\in I : aR_i a' \ \forall a' \in A\}} p_i.$$

(iii) If $\bigcap_{i\in I} A_i \neq \varnothing$, then for all $a \in \bigcap_{i\in I} A_i$,

$$x(s,a) = \sum_{\{i\in I : aR_i a' \ \forall a' \in \bigcap_{i\in I} A_i\}} p_i.$$

In words, this game form offers each agent i the opportunity to provide a complete ranking of outcomes R_i and also a set A_i of "acceptable" alternatives. If there is at least one common element among the sets of acceptable alternatives for all agents, then the mechanism implements random dictatorship (with the preferences described by the R_i), but with the restriction that the dictator can only choose an outcome from the unanimously acceptable alternatives. Otherwise, the mechanism reverts to random dictatorship (with outcomes determined by the highest ranked elements of the R_i). We refer to this game form as *p-random dictatorship with compromise* because it offers agents the opportunity to replace the outcome of p-random dictatorship with a compromise on a mutually acceptable alternative.[18]

It is elementary to verify that a strategy of player i, such that for some type t_i, we have $b(u_i(t_i)) \notin A_i$, is weakly dominated by the same strategy in which A_i is replaced by $A_i \cup \{b(u_i(t_i))\}$. Moreover, any strategy of some player i, such that for some type t_i, we have that R_i is not type t_i's true preference over A_i as described by $u_i(t_i)$, is weakly dominated by a strategy such that A_i is left unchanged, but R_i is replaced by a preference ordering that reflects t_i's true preference over A_i. Preferences that player i indicates for alternatives $A \setminus A_i$ are irrelevant for the outcome of the game. These considerations motivate us to restrict attention to "truthful strategies," which we define to be strategies such that $b(u_i(t_i)) \in A_i$ and such that R_i is the true preference according to $u_i(t_i)$ for all types t_i. Note that we have ruled out some, but not necessarily all, weakly dominated

[18]This game form was inspired by *approval voting* (see Brams and Fishburn 2007), which, like our game form, allows voters to indicate acceptable alternatives. However, in approval voting the alternative that the largest number of agents regards as acceptable is selected, whereas our game form requires unanimity. Moreover, our game form uses random dictatorship as a fall-back, whereas approval voting does not have any such fall-back. When p is the uniform distribution, the game form that we consider is closely related to the full consensus or random ballot fall-back game form that Heitzig and Simmons (2012) introduced. Heitzig and Simmons require the sets A_i to be singletons.

strategies. In any case, it seems eminently plausible that all players will choose truthful strategies.

In a Bayesian equilibrium for all type spaces in which all players choose truthful strategies, any type's interim expected utility is not smaller than the interim expected utility from p-random dictatorship. This is because a type can always force an outcome that gives at least as high interim expected utility as p-random dictatorship by choosing the truthful strategy for which $A_i = \{b(u_i(t_i))\}$. Note also that it is a consistent Bayesian equilibrium for all type spaces that all players choose this strategy for all types.

We now show that *p-random dictatorship with compromise* also has a Bayesian equilibrium for all type spaces in which all players choose truthful strategies and that interim Pareto dominates random dictatorship. We further show that this equilibrium respects positive and negative unanimity. The latter observation clarifies that our result is indeed a consequence of weakening the belief independence requirement and not of weakening any other property listed in Proposition 1.

PROPOSITION 2. *For every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$ and $p_i < 1$ for all $i \in I$, p-random dictatorship with compromise has a consistent Bayesian equilibrium for all type spaces σ^* that interim Pareto dominates p-random dictatorship and that satisfies positive and negative unanimity.*

PROOF. We construct the equilibrium σ^* inductively. We begin by considering type spaces $\mathcal{T}$, where for every $i \in I$, the set T_i has exactly one element. In such type spaces, for every $i \in I$, it is common knowledge among the agents that agent i has utility function $u_i(t_i)$. We distinguish two cases. The first is that there is some alternative $a \in A$ such that for all $i \in I$, we have

$$u_i(t_i, a) > \sum_{j\in I} p_j u_i(t_i, b(u_j(t_j))). \tag{1}$$

Observe that the assumption $p_i < 1$ for all $i \in I$ implies that some such type spaces exist. For such type spaces, the strategies are

$$\sigma_i^*(\mathcal{T}, t_i) = (\{b(u_i(t_i)), a\}, R_i) \tag{2}$$

for all $i \in I$, where R_i is agent i's true preference and where a is some alternative for which (1) holds.[19] These strategies obviously constitute a Nash equilibrium of the complete information game in which agents' preferences are common knowledge. Note that the outcome a then results and that this outcome strictly Pareto-dominates the outcome of random dictatorship.

For all other type spaces with just a single element for each player, the strategies are

$$\sigma_i^*(\mathcal{T}, t_i) = (\{b(u_i(t_i))\}, R_i)$$

for all $i \in I$, where R_i is again agent i's true preference. We noted earlier that these strategies constitute a Nash equilibrium and that the outcome is the same as under p-random dictatorship.

[19] Note that a must be the same for all players.

Now suppose we had constructed the equilibrium for all type spaces $\mathcal{T}$ in which the sum over i of the numbers of elements of the sets T_i is at most k, that is, $\sum_{i\in I}|T_i| \leq k$. We extend the construction to all type spaces $\mathcal{T}$ in which this sum is $k+1$. Fix some particular such type space $\mathcal{T}$. Consider all type spaces $\widetilde{\mathcal{T}}$ that are contained in $\mathcal{T}$, that is, for which conditions (i) and (ii) of Definition 10 hold, and that are not equal to $\mathcal{T}$. For such type spaces, we define for every $i \in I$ and every $t_i \in \widetilde{T}_i$,

$$\sigma_i^*(\mathcal{T}, t_i) = \sigma_i^*(\widetilde{\mathcal{T}}, t_i).$$

By the inductive hypothesis, the right hand side of this equation has already been defined. Observe that the right hand side does not depend on the particular choice of $\widetilde{\mathcal{T}}$. If a type t_i of player i is contained in player i's type set in two different type spaces $\widetilde{\mathcal{T}}$ and $\widehat{\mathcal{T}}$ that are contained in $\mathcal{T}$ in the sense of Definition 10, then the intersection of these type spaces is also a type space and, by consistency, the same strategy is assigned to type t_i in $\widetilde{\mathcal{T}}$ and in $\widehat{\mathcal{T}}$.

If the previous step defines the equilibrium strategy for all types in $\mathcal{T}$, then the inductive step is completed. Otherwise, it remains to define strategies for types t_i that are not contained in any type set of a type space that is a subspace of $\mathcal{T}$. We consider the strategic game in which each such type is a separate player and expected utilities are calculated, keeping the strategies of types that have already been dealt with in the previous paragraph fixed, and using each type's subjective beliefs to calculate that type's expected payoff. We restrict attention to truthful strategies and strategies such that $w(u_i(t_i)) \notin A_i$. This strategic game has a Nash equilibrium in mixed strategies, and this Nash equilibrium is also a Nash equilibrium of the game with unrestricted strategy spaces. For each type t_i that still has to be dealt with, we define the strategy $\sigma_i^*(\mathcal{T}, t_i)$ to be type t_i's equilibrium strategy.

By construction, these strategies satisfy the consistency requirement. Also, they are by construction interim Bayesian equilibria: For types in type sets that correspond to a smaller type space, the Bayesian equilibrium property carries over from the smaller type space. For all other types, their choices maximize expected utility by construction.

The equilibrium that we have constructed interim Pareto dominates random dictatorship. First, we note that when all players choose truthful strategies, no type can have lower expected utility than under random dictatorship. This is because each type can guarantee themselves an outcome that is at least as good as the random dictatorship outcome by choosing $A_i = \{b(u_i(t_i))\}$. Second, each type's expected utility is increased on type spaces in which each player's type set has just a single element and for which inequality (1) holds.

The equilibrium that we have constructed satisfies positive unanimity because all players include their most preferred alternative in the set A_i. If all players have the same most preferred alternative, the sets A_i will have a nonempty intersection and the random dictator will select the alternative that is most preferred by everyone. The equilibrium also satisfies negative unanimity. We have assumed that no player includes his/her least preferred alternative in the set A_i. Therefore, independent of whether these sets have a nonempty intersection or not, the random dictator will not select the agents' least preferred alternative if they all have the same least preferred alternative. □

In the above proof, we could have replaced the second step of our induction by an appeal to Propositions 2 and 3 in Yildiz (2011). This is because the second step extends the equilibrium construction from a set of small type spaces to a set of larger type spaces, making sure that the equilibrium for the larger type space is consistent, and Yildiz's results show the possibility of such an extension for a class of invariance requirements that includes consistency, using a more general version of the argument that we have used above. To make this paper self-contained, we have given a complete proof of Proposition 2.

It is obvious that the proof of Proposition 2 also proves the following result.

REMARK 1. If $p_i < 1$ for all $i \in I$, then every Bayesian equilibrium for all type spaces of p-random dictatorship with compromise, in which all players choose truthful strategies, and in which players choose strategies of the form (2) in type spaces in which each player's type set has just a single element and for which inequality (1) holds,[20] interim Pareto dominates p-random dictatorship.

Note that the set of truthful strategies includes, in particular, the set of all non-weakly-dominated strategies. Therefore, Remark 1 applies to all equilibria in non-weakly-dominated strategies. As we explained in the Introduction, the importance of this result is that it shows that p-random dictatorship with compromise interim Pareto dominates p-random dictatorship not just in the sense of mechanism design (where we only have to find one equilibrium with the desired properties), but also in the sense of implementation (where all equilibria—or, as in our case, all that satisfy some refinement—are considered). Observe, incidentally, that for the result of Remark 1, the Bayesian equilibrium for all type spaces does not need to be consistent.

We do not know whether Proposition 2 remains true if we include a universal type space, such as Mertens and Zamir (1985) or Sadzik (2011) universal type space, in the set of type spaces that we consider or if we consider equilibria on the universal type space alone. No Bayesian equilibrium on any type space in truthful strategies can make any player worse off at the interim level than random dictatorship, because each agent can always unilaterally enforce a return to random dictatorship. This step presents no difficulty. Proving the existence of *some* consistent Bayesian equilibrium is not problematic either. First, we can ignore the consistency requirement and all type spaces except the universal type space, because (as we noted in Section 5) we can derive from any Bayesian equilibrium on a universal type space a consistent Bayesian equilibrium for all type spaces. Second, it is immediate that it is a Bayesian equilibrium of p-random dictatorship with compromise on the universal type space that all agents choose as if there were no possibility to compromise. The problem is that this equilibrium is equivalent to p-random dictatorship. There is no type vector for which it is interim Pareto superior to p-random dictatorship, as required by interim Pareto dominance.

[20] In words, this condition says that if utility functions are common knowledge and some alternative Pareto dominates random dictatorship, then players pick their preferred alternative and some alternative—the same for all players—that Pareto dominates random dictatorship as their set of acceptable alternatives.

To see the difficulty in finding a suitable equilibrium on the universal type space, consider what would happen if we tried to adapt the argument in the proof of Proposition 2 to a universal type space. We would again begin by considering subsets of the universal type space in which payoffs are common knowledge and we would define equilibrium for such subsets as in the proof of Proposition 2. We would then seek to extend the equilibrium, defining strategies for all types in the universal type space. The difficulty is that we do not know whether this extension is possible. When such an extension is possible for every Bayesian equilibrium, Friedenberg and Meier (2010) say that the original game, the restricted type space, and the larger type space have the "equilibrium extension property." Friedenberg and Meier give an example where the larger type space is Mertens and Zamir's universal type space and where the extension property fails to hold for some equilibrium. They also provide sufficient conditions for the equilibrium extension property to be satisfied, but it is immediate that these conditions do not hold in our example.

7. No game form ex post Pareto dominates random dictatorship

In this section we show that the result of the previous section does not hold if utilities are evaluated ex post. The following proposition shows that in fact no mechanism ex post Pareto dominates p-random dictatorship.

Proposition 3. *For every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$, there is no game form G that has a consistent equilibrium for all type spaces σ^* that ex post Pareto dominates p-random dictatorship.*

Proof. The proof is indirect. Suppose for some $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$, there were a game form G and a consistent Bayesian equilibrium of G for all type spaces σ^* that ex post Pareto dominated p-random dictatorship. For the outcome that results from G and σ^* to be different from p-random dictatorship, there must be some $\widehat{\mathcal{T}} \in \Upsilon$, $\hat{t} \in \widehat{T}$, and $\hat{a} \in A$ such that

$$\sum_{s\in S} x(s,\hat{a}) \cdot \prod_{i\in I} \sigma_i^*(\widehat{\mathcal{T}}, \hat{t}_i, s_i) < \sum_{\{i\in I : b(u_i(\hat{t}_i))=\hat{a}\}} p_i. \tag{3}$$

That is, alternative $\hat{a}$ is chosen with a probability that is strictly smaller than the probability with which it is chosen under random dictatorship. Let $\widehat{I}$ be the set $\{i \in I : b(u_i(\hat{t}_i)) = \hat{a}\}$. Notice that we must have $\varnothing \neq \widehat{I} \neq I$. If $\widehat{I} = \varnothing$, the right hand side of (3) would be zero. If $\widehat{I} = I$, then G and σ^* would be ex post Pareto worse than random dictatorship at $\hat{t}$. To complete the proof, we construct a new type space $\widetilde{\mathcal{T}}$ and infer from (3) that in this type space there is a type vector such that the types of all players in $\widehat{I}$ strictly prefer the outcome of p-random dictatorship conditional on this type vector to the outcome in G resulting from the equilibrium σ^* conditional on this type vector. Therefore, G and σ^* do not ex post Pareto dominate p-random dictatorship.

The type sets in $\widetilde{\mathcal{T}}$ are given by $\widetilde{T}_i = \widehat{T}_i$ for all $i \in \widehat{I}$ and $\widetilde{T}_i = \widehat{T}_i \cup \{\tilde{t}_i\}$ for all $i \notin \widehat{I}$. For all $i \in I$, the types in $\widehat{T}_i$ have the same utility functions and beliefs in $\widetilde{\mathcal{T}}$ as in $\widehat{\mathcal{T}}$. For all $i \notin \widehat{I}$,

type $\tilde{t}_i$'s beliefs are given by

$$\pi_i(\tilde{t}_i)[((\hat{t}_j)_{j\in\widehat{I}}, (\tilde{t}_j)_{\substack{j\notin\widehat{I}\\ j<i}}, (\hat{t}_j)_{\substack{j\notin\widehat{I}\\ j>i}})] = 1$$

and type $\tilde{t}_i$'s utility function is

$$\tilde{u}_i(\tilde{t}_i, a) = \begin{cases} 1 & \text{if } a = \tilde{a} \\ 1-\varepsilon_a & \text{if } a \notin \{\hat{a}, \tilde{a}\} \\ 0 & \text{if } a = \hat{a}, \end{cases}$$

where $\tilde{a}$ denotes the second most preferred alternative of some player k's type $\hat{t}_k$, where $k \in \widehat{I}$. We assume that $0 < \varepsilon_a < \bar{\varepsilon}$ for all $a \notin \{\hat{a}, \tilde{a}\}$ for some $\bar{\varepsilon} \in (0,1)$, and that $a, a' \notin \{\hat{a}, \tilde{a}\}$ and $a \neq a'$ implies $\varepsilon_a \neq \varepsilon_{a'}$. This assumption ensures that the utility functions satisfy the condition of no indifferences. Moreover, by letting $\bar{\varepsilon}$ tend to zero, we can ensure that all ε_a tend to zero, which is the case that we shall focus on.

We now show that for $\bar{\varepsilon}$ sufficiently small at type vector $((\hat{t}_i)_{i\in\widehat{I}}, (\tilde{t}_i)_{i\notin\widehat{I}})$, the alternatives other than $\hat{a}$ are in equilibrium σ^* chosen with a probability larger than $1-\sum_{i\in\widehat{I}} p_i$. Note that the proof of Proposition 3 is concluded once this assertion is established. This is because random dictatorship gives for some $k \in \widehat{I}$ player k's type $\hat{t}_k$, his/her top alternative $\hat{a}$ with probability $\sum_{i\in\widehat{I}} p_i$ and type $\hat{t}_k$'s second most preferred alternative $\tilde{a}$ with probability $1-\sum_{i\in\widehat{I}} p_i$. By contrast, G and σ^* yield $\hat{a}$ with probability less than $\sum_{i\in\widehat{I}} p_i$ and some other alternative, not necessarily type $\hat{t}_k$'s second most preferred alternative, with a probability larger than $1-\sum_{i\in\widehat{I}} p_i$. Therefore, type $\hat{t}_k$ strictly prefers random dictatorship.

Consider the player $i \notin \widehat{I}$ for whom i is smallest. We denote this player by $i1$. This player, when type $\tilde{t}_{i1}$, expects with probability 1 that the other players' type vector is $\hat{t}_{-i1}$. Because σ^* is consistent, type $\tilde{t}_{i1}$ expects the types $\hat{t}_{-i1}$ to choose the same in $\widetilde{T}$ as in $\widehat{T}$. By the assumption of the indirect proof, type $\hat{t}_{i1}$ has a strategy available that yields alternatives other than $\hat{a}$ with probability of more than $1-\sum_{i\in\widehat{I}} p_i$. Type $\tilde{t}_{i1}$ will not necessarily choose the same strategy as type $\hat{t}_{i1}$. But, for small enough $\bar{\varepsilon}$, only a strategy that yields an alternative other than $\hat{a}$ with some probability $\tilde{p} > 1-\sum_{i\in\widehat{I}} p_i$ can be optimal. Choosing such a strategy yields for type $\tilde{t}_{i1}$ expected payoff greater than $\tilde{p}(1-\bar{\varepsilon})$, whereas any other pure strategy yields a payoff that is no more than $1-\sum_{i\in\widehat{I}} p_i < \tilde{p}$. For small enough $\bar{\varepsilon}$, the former expected payoff is larger than the latter.

Now consider the player $i \notin \widehat{I}$ for whom i is second smallest. We denote this player by $i2$. This player, when type $\tilde{t}_{i2}$, expects with probability 1 the other players' types to be $\hat{t}_{-i2}$ except for player $i1$, whom $i2$ expects with probability 1 to be type $\tilde{t}_{i1}$. By the step of the previous paragraph, if $\tilde{t}_{i2}$ chose the same strategy as $\hat{t}_{i2}$ does in equilibrium, $\tilde{t}_{i2}$ would expect an outcome other than $\hat{a}$ with probability larger than $1-\sum_{i\in\widehat{I}} p_i$. He might choose in equilibrium some other strategy, but, for small enough $\bar{\varepsilon}$, he will never make a choice that yields an outcome other than $\hat{a}$ with a probability that is not larger than $1-\sum_{i\in\widehat{I}} p_i$.

The step of the previous paragraph can be iterated until we arrive at the player $i \notin \widehat{I}$ for whom i is largest. We denote this player by $i(m)$. This player expects the other players

to be of type $\tilde{t}_{-(i(m))}$ except for types $i \in \widehat{I}$, whom this player expects to be of type $\hat{t}_i$. By the same argument used in the previous two paragraphs, type $\tilde{t}_{i(m)}$ chooses in equilibrium a strategy that he expects to yield an outcome other than $\hat{a}$ with probability larger than $1 - \sum_{i \in \widehat{I}} p_i$. But at type vector $((\hat{t}_i)_{i \in \widehat{I}}, (\tilde{t}_i)_{i \notin \widehat{I}})$, this type has correct expectations and, therefore, at this type vector, the equilibrium strategies do indeed yield an outcome other than $\hat{a}$ with probability larger than $1 - \sum_{i \in \widehat{I}} p_i$. As explained above, this concludes the proof. □

If our solution concept were Bayesian equilibrium without refinement rather than consistent equilibrium, Proposition 3 would be false. The p-random dictatorship with compromise has the following Bayesian equilibrium that ex post Pareto dominates p-random dictatorship. For the common knowledge type spaces referred to in the first paragraph of the proof of Proposition 2, agents play the strategies described in that paragraph. For all other type spaces, they ignore the possibility of compromise. We do not find this equilibrium plausible. It implies that agents' choices in the case that preferences are common knowledge depend on whether that common knowledge is represented by a single element type space or by a larger type space that includes as a belief-closed subset the same single element type space. As we argued in Section 5, this means that agents' choices not only depend on agents' beliefs, but also on the way in which the modeler represents those beliefs, which, to us, does not seem to make sense.

Proposition 3 would remain true if we included a universal type space in the set of all type spaces that we consider, and it would also remain true if the universal type space were the only type space that we considered. Indeed, one could then drop the consistency requirement and focus on the universal type space, even if other type spaces were included in the model. The argument in the proof of Proposition 3 would go through without alteration.

8. Conclusion

Gibbard and Satterthwaite's impossibility theorem, and Hylland's version of this theorem in a setting with stochastic outcomes are central results of voting theory. We have argued that the insistence of these theorems on belief-independent strategy choices may be overly restrictive if a mechanism designer is concerned with Pareto improvements. Such a mechanism designer can find voting schemes that are superior to random dictatorship if agents' choices are allowed to depend on their beliefs. Whatever those beliefs are, the outcomes will be at least as good as under random dictatorship and sometimes better. Such an improvement is only possible if agents' subjective beliefs are accepted and an interim perspective is adopted. From an ex post perspective, such unambiguous improvements are not possible.

An important problem left open by our paper is the characterization of voting rules that are not dominated in one of the senses considered here. One can take a mechanism design or an implementation approach to this question, depending on whether one considers just one or all consistent Bayesian equilibria on all type spaces of a given game

form. In Smith (2010), the analogous question is investigated for public good mechanisms, using a mechanism design approach. Smith proves for one particular mechanism that it is not dominated. Smith's work shows the subtleties involved in such proofs. We leave the question as applied to voting rules for future research.

References

Azrieli, Yaron and Semin Kim (forthcoming), "Pareto efficiency and weighted majority rules." *International Economic Review*. [344]

Barberà, Salvador (2010), "Strategy-proof social choice." In *Handbook of Social Choice and Welfare*, Vol. 2 (Kenneth J. Arrow, Amartya K. Sen, and Kotaro Suzumura, eds.), North-Holland, Amsterdam. [340]

Bergemann, Dirk and Stephen Morris (2003), "Robust mechanism design." Discussion Paper 1421, Cowles Foundation. [343]

Bergemann, Dirk and Stephen Morris (2005), "Robust mechanism design." *Econometrica*, 73, 1771–1813. [342, 343]

Bergemann, Dirk and Stephen Morris (2011), "Robust implementation in general mechanisms." *Games and Economic Behavior*, 71, 261–281. [344]

Bergemann, Dirk and Stephen Morris (2012), "Robust mechanism design: An introduction." In *Robust Mechanism Design* (Dirk Bergemann and Stephen Morris, eds.), World Scientific Publishing, Singapore. [342]

Blin, Jean-Marie and Mark Satterthwaite (1977), "On preferences, beliefs, and manipulation within voting situations." *Econometrica*, 45, 881–888. [340]

Börgers, Tilman (1991), "Undominated strategies and coordination in normalform games." *Social Choice and Welfare*, 8, 65–78. [344]

Börgers, Tilman and Peter Postl (2009), "Efficient compromising." *Journal of Economic Theory*, 144, 2057–2076. [344]

Börgers, Tilman and Doug Smith (2012), "Robustly ranking mechanisms." *American Economic Review: Papers and Proceedings*, 102, 325–329. [344]

Brams, Steven and Peter Fishburn (2007), *Approval Voting*, second edition. Springer, Heidelberg. [352]

Chung, Kim-Sau and Jeff Ely (2007), "Foundations of dominant strategy mechanisms." *Review of Economic Studies*, 74, 447–476. [343]

Dutta, Bhaskar, Hans Peters, and Arunava Sen (2007), "Strategy-proof cardinal decision schemes." *Social Choice and Welfare*, 28, 163–179. [341, 349]

Dutta, Bhaskar, Hans Peters, and Arunava Sen (2008), "Strategy-proof cardinal decision schemes." *Social Choice and Welfare*, 30, 701–702 (erratum). [341, 349, 350]

Ely, Jeffrey and Marcin Pęski (2006), "Hierarchies of belief and interim rationalizability." *Theoretical Economics*, 1, 19–65. [346]

Friedenberg, Amanda and Martin Meier (2010), "The context of the game." Unpublished paper, Arizona State University. [351, 356]

Gibbard, Allan (1973), "Manipulation of voting schemes: A general result." *Econometrica*, 41, 587–601. [339]

Heitzig, Jobst and Forrest Simmons (2012), "Some chance for consensus: Voting methods for which consensus is an equilibrium." *Social Choice and Welfare*, 38, 43–57. [344, 352]

Holmström, Bengt and Roger B. Myerson (1983), "Efficient and durable decision rules with incomplete information." *Econometrica*, 51, 1799–1819. [341]

Hylland, Aanund (1980), "Strategy proofness of voting procedures with lotteries as outcomes and infinite sets of strategies." Unpublished paper, University of Oslo. [341, 349]

Jackson, Matthew O. (2001), "A crash course in implementation theory." *Social Choice and Welfare*, 18, 655–708. [341]

Mertens, Jean-François and Shmuel Zamir (1985), "Formulation of Bayesian analysis for games with incomplete information." *International Journal of Game Theory*, 14, 1–29. [342, 346, 355]

Nandeibam, Shasikanta (2013), "The structure of decision schemes with cardinal preferences." *Review of Economic Design*, 17, 205–238. [341]

Sadzik, Tomasz (2011), "Beliefs revealed in Bayesian-Nash equilibrium." Unpublished paper, New York University. [346, 355]

Satterthwaite, Mark A. (1975), "Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions." *Journal of Economic Theory*, 10, 187–217. [339]

Schmitz, Patrick W. and Thomas Tröger (2012), "The (sub-)optimality of the majority rule." *Games and Economic Behavior*, 74, 651–665. [344]

Smith, Doug (2010), "A prior free efficiency comparison of mechanisms for the public goods problem." Unpublished paper, University of Michigan, Ann Arbor. [343, 359]

Yamashita, Takuro (2012), "A necessary condition for implementation in undominated strategies, with applications to robustly optimal trading mechanisms." Unpublished paper, Toulouse School of Economics. [344]

Yildiz, Muhamet (2011), "Invariance to representation of information." Unpublished paper, MIT. [351, 355]

Zeckhauser, Richard (1973), "Voting systems, honest preferences, and Pareto optimality." *American Political Science Review*, 67, 934–946. [344]

Submitted 2011-11-3. Final version accepted 2013-1-18. Available online 2013-1-18.