Economics and Philosophy, **23** (2007) 269–300
doi:10.1017/S0266267107001496

STRATEGY-PROOF JUDGMENT AGGREGATION*

Franz Dietrich
University of Maastricht

Christian List
London School of Economics

Which rules for aggregating judgments on logically connected propositions are manipulable and which not? In this paper, we introduce a preference-free concept of non-manipulability and contrast it with a preference-theoretic concept of strategy-proofness. We characterize all non-manipulable and all strategy-proof judgment aggregation rules and prove an impossibility theorem similar to the Gibbard–Satterthwaite theorem. We also discuss weaker forms of non-manipulability and strategy-proofness. Comparing two frequently discussed aggregation rules, we show that "conclusion-based voting" is less vulnerable to manipulation than "premise-based voting", which is strategy-proof only for "reason-oriented" individuals. Surprisingly, for "outcome-oriented" individuals, the two rules are strategically equivalent, generating identical judgments in equilibrium. Our results introduce game-theoretic considerations into judgment aggregation and have implications for debates on deliberative democracy.

* F. Dietrich, Dept. of Quant. Econ., Univ. of Maastricht, P.O. Box 616, 6200 MD Maastricht, NL. C. List, Dept. of Govt., LSE, London WC2A 2AE, UK. This paper was presented at the University of Konstanz (6/2004), the Social Choice and Welfare Conference in Osaka (7/2004), the London School of Economics (10/2004), Université de Caen (11/2004), the University of East Anglia (1/2005), Northwestern University (5/2005), the 2005 SAET Conference in Vigo (6/2005), the University of Hamburg (10/2005), IHPST, Paris (1/2006). We thank the participants at these occasions, the anonymous referees of this paper and the editor, Bertil Tungodden, for comments.

1. INTRODUCTION

How can a group of individuals aggregate their individual judgments (beliefs, opinions) on some logically connected propositions into collective judgments on these propositions? In particular, how can a group do this under conditions of pluralism, i.e., when individuals disagree on the propositions in question? This problem – *judgment aggregation* – is discussed in a growing literature in philosophy, economics and political science and generalizes earlier problems of social choice, notably preference aggregation in the Condorcet–Arrow tradition.[1] The problem arises in many different decision-making bodies, ranging from legislative committees and multi-member courts to expert advisory panels and monetary policy committees of a central bank.

Judgment aggregation is often illustrated by a paradox: the *discursive* (or *doctrinal*) *paradox* (Kornhauser and Sager 1986; Pettit 2001; Brennan 2001). To illustrate, suppose a university committee responsible for a tenure decision has to make collective judgments on three propositions:[2]

a: The candidate is good at teaching.
b: The candidate is good at research.
c: The candidate deserves tenure.

According to the university's rules, c (the "conclusion") is true if and only if a and b (the "premises") are both true, formally $c \leftrightarrow (a \wedge b)$ (the "connection rule"). Suppose the committee has three members with judgments as shown in Table 1.

If the committee takes a majority vote on each proposition, then a and b are each accepted and yet c is rejected (each by two thirds), despite the (unanimous) acceptance of $c \leftrightarrow (a \wedge b)$. The discursive paradox shows that judgment aggregation by propositionwise majority voting may lead to inconsistent collective judgments, just as Condorcet's paradox shows that preference aggregation by pairwise majority voting may lead to intransitive collective preferences.

In response to the discursive paradox, two aggregation rules have been proposed to avoid such inconsistencies (e.g., Pettit 2001; Chapman 1998, 2002; Bovens and Rabinowicz 2006). Under *premise-based voting*, majority votes are taken on a and b (the premises), but not on c (the conclusion), and the collective judgment on c is derived using the connection rule $c \leftrightarrow (a \wedge b)$: in Table 1, a, b and c are all accepted. Premise-based voting captures the deliberative democratic idea that collective decisions on outcomes should

[1] Preference aggregation becomes a case of judgment aggregation by expressing preference relations as sets of binary ranking propositions in predicate logic (List and Pettit 2004; Dietrich and List 2007a).

[2] This example is due to Bovens and Rabinowicz (2006).

TABLE 1. The discursive paradox

	a	b	$c \leftrightarrow (a \wedge b)$	c
Individual 1	Yes	Yes	Yes	Yes
Individual 2	Yes	No	Yes	No
Individual 3	No	Yes	Yes	No
Majority	Yes	Yes	Yes	No

be made on the basis of collectively decided reasons. Here reasoning is "collectivized", as Pettit (2001) describes it. Under *conclusion-based voting*, a majority vote is taken only on c, and no collective judgments are made on a or b: in Table 1, c is rejected and other propositions are left undecided. Conclusion-based voting captures the minimal liberal idea that collective decisions should be made only on (practical) outcomes and that the reasons behind such decisions should remain private. Here collective decisions are "incompletely theorized" in Sunstein's (1994) terms. (For a comparison between minimal liberal and comprehensive deliberative approaches to decision making, see List 2006.)

Abstracting from the discursive dilemma, List and Pettit (2002, 2004) have formalized judgment aggregation and proved that no judgment aggregation rule ensuring consistency can satisfy some conditions inspired by Arrow's conditions on preference aggregation. This impossibility result has been strengthened and extended by Pauly and van Hees (2006; see also van Hees 2007), Dietrich (2006), Gärdenfors (2006) and Dietrich and List (2007a, 2007b). Drawing on the model of "property spaces", Nehring and Puppe (2002, 2005) have offered the first characterizations of agendas of propositions for which impossibility results hold (for a subsequent contribution, see Dokow and Holzman 2005). Possibility results have been obtained by List (2003, 2004), Pigozzi (2006) and Osherson and Vardi (forthcoming). Dietrich (2007) has developed an extension of the judgment aggregation model to richer logical languages for expressing propositions, which we use in this paper. Related bodies of literature include those on abstract aggregation theory (Wilson 1975)[3] and on belief merging in computer science (Konieczny and Pino-Perez 2002).

[3] Wilson's (1975) aggregation problem, where a group has to form yes/no views on several issues based on individual views on them (subject to feasibility constraints), can be represented in judgment aggregation. Unlike judgment aggregation, Wilson's model cannot fully generally represent logical entailment: its primitive is a consistency (feasibility) notion, from which an entailment relation can be retrieved only for certain logical languages (Dietrich 2007).

But one important question has received little attention in the literature on judgment aggregation: Which aggregation rules are manipulable by strategic voting and which are strategy-proof? The answer is not obvious, as strategy-proofness in the familiar sense in economics is a preference-theoretic concept and preferences are not primitives of judgment aggregation models. Yet the question matters for the design and implementation of an aggregation rule in a collective decision-making body such as in the examples above. Ideally, we would like to find aggregation rules that lead individuals to reveal their judgments truthfully. Indeed, if an aggregation rule captures the normatively desirable functional relation between individual and collective judgments, then truthful revelation of these individual judgments (which are typically private information) is crucial for the (direct) implementation of that functional relation.[4]

In this paper, we address this question. We first introduce a simple condition of non-manipulability and characterize the class of non-manipulable judgment aggregation rules. We then show that, under certain motivational assumptions about individuals, our condition is equivalent to a game-theoretic strategy-proofness condition similar to the one introduced by Gibbard (1973) and Satterthwaite (1975) for preference aggregation.[5] Our characterization of non-manipulable aggregation rules then yields a characterization of strategy-proof aggregation rules. The relevant motivational assumptions hold if agents want the group to make collective judgments that match their own individual judgments (e.g., want the group to make judgments that match what they consider the truth). In many other cases, such as that of "reason-oriented" individuals (as defined in Section 5), non-manipulability and strategy-proofness may come significantly apart.

By introducing both a non-game-theoretic condition of non-manipulability and a game-theoretic condition of strategy-proofness, we are able to distinguish between *opportunities* for manipulation (which depend only on the aggregation rule in question) and *incentives* for manipulation (which depend also on the motivations of the decision-makers).

We prove that, for a general class of aggregation problems including the tenure example above, there exists no non-manipulable judgment aggregation rule satisfying universal domain and some other mild

[4] A functional relation between individual and collective judgments could be deemed normatively desirable for a variety of reasons, such as epistemic or democratic legitimacy goals. The axiomatic approach to social choice theory translates these goals into formal requirements on aggregation.

[5] Our definition of strategy-proofness in judgment aggregation draws on List (2002b, 2004), where sufficient conditions for strategy-proofness in (sequential) judgment aggregation are given.

conditions, an impossibility result similar to the Gibbard–Satterthwaite theorem on preference aggregation. Subsequently, we identify various ways to avoid the impossibility result. We also show that our default conditions of non-manipulability and strategy-proofness fall into general families of conditions and discuss other conditions in these families. In the case of strategy-proofness, these conditions correspond to different motivational assumptions about the decision makers. In the tenure example, conclusion-based voting is strategy-proof in a strong sense, but produces no collective judgments on the premises. Premise-based voting satisfies only the weaker condition of strategy-proofness for "reason-oriented" individuals. Surprisingly, although premise- and conclusion-based voting are regarded in the literature as two diametrically opposed aggregation rules, they are strategically equivalent if individuals are "outcome-oriented", generating identical judgments in equilibrium. Our results not only introduce game-theoretic considerations into the theory of judgment aggregation, but they are also relevant to debates on democratic theory as premise-based voting has been advocated, and conclusion-based voting rejected, by proponents of deliberative democracy (Pettit 2001).

There is, of course, a related literature on manipulability and strategy-proofness in preference aggregation, following Gibbard's and Satterthwaite's classic contributions (e.g., Taylor 2002, 2005; Saporiti and Thomé 2005). An important branch of this literature, from which several corollaries for judgment aggregation can be derived, has considered preference aggregation over options that are vectors of binary properties (Barberà et al. 1993, 1997; Nehring and Puppe 2002). A parallel to judgment aggregation can be drawn by identifying propositions with properties; a disanalogy lies in the structure of the informational input to the aggregation rule. While judgment aggregation rules collect a single judgment set from each individual (expressed in a possibly rich logical language), preference aggregation rules collect an entire preference ordering over vectors of properties. Whether or not an individual's most-preferred vector of properties (in preference aggregation) can be identified with her judgment set (in judgment aggregation) depends precisely on the motivational assumptions we make about this individual.

Another important related literature is that on the paradox of multiple elections (Brams et al. 1997, 1998; Kelly 1989). Here a group also aggregates individual votes on multiple propositions, and the winning combination can be one that no voter individually endorses. However, given the different focus of that work, the propositions in question are not explicitly modelled as logically interconnected as in our present model of judgment aggregation. The formal proofs of all the results reported in the main text are given in the Appendix.

2. THE BASIC MODEL

We consider a group of individuals $N = \{1, 2, \ldots, n\}$, where $n \geq 2$.[6] The group has to make collective judgments on logically connected propositions.

2.1 Representing propositions in formal logic

Propositions are represented in a *logical language*, defined by two components:

- a non-empty set $\mathbf{L}$ of formal expressions representing *propositions*; the language has a negation symbol $\neg$ ("not"), where for each proposition p in $\mathbf{L}$, its negation $\neg p$ is also contained in $\mathbf{L}$.
- an *entailment relation* $\models$, where, for each set of propositions $A \subseteq \mathbf{L}$ and each proposition $p \in \mathbf{L}$, $A \models p$ is read as "A logically entails p".[7]

We call a set of propositions $A \subseteq \mathbf{L}$ *inconsistent* if $A \models p$ and $A \models \neg p$ for some $p \in \mathbf{L}$, and *consistent* otherwise. We require the logical language to have certain minimal properties (Dietrich 2007; Dietrich and List 2007a).[8]

The most familiar logical language is (*classical*) *propositional logic*, containing a given set of *atomic* propositions $a, b, c, \ldots$, such as the propositions about the candidate's teaching, research and tenure in the example above, and *compound* propositions with the logical connectives $\neg$ ("not"), $\wedge$ ("and"), $\vee$ ("or"), $\rightarrow$ ("if-then"), $\leftrightarrow$ ("if and only if"), such as the connection rule $c \leftrightarrow (a \wedge b)$ in the tenure example.[9] Examples of valid logical entailments in propositional logic are $\{a, \{a \rightarrow b\} \models b$ ("modus ponens"), $a \rightarrow b, \neg b\} \models \neg a$ ("modus tollens"), whereas the entailment $\{a \vee b\} \models a$ is not valid. Examples of consistent sets are $\{a, a \vee$

[6] Although no discursive paradox arises for $n = 2$, our results below still hold: Under Theorem 2's other conditions, non-manipulability requires a dictatorship of one of the two individuals. The unanimity rule, while also non-manipulable, violates completeness of collective judgments.

[7] $\models$ can be interpreted either as semantic entailment or as syntactic derivability (usually denoted $\vdash$). The two interpretations give rise to semantic or syntactic notions of rationality, respectively.

[8] L1 (self-entailment): For all $p \in \mathbf{L}$, $\{p\} \models p$. L2 (monotonicity): For all $p \in \mathbf{L}$ and $A \subseteq B \subseteq \mathbf{L}$, if $A \models p$ then $B \models p$. L3 (completability): $\emptyset$ is consistent, and each consistent set $A \subseteq \boldsymbol{L}$ has a consistent superset $B \subseteq \mathbf{L}$ containing a member of each pair $p, \neg p \in \mathbf{L}$. L1–L3 are jointly equivalent to three conditions on the consistency notion: each pair $\{p, \neg p\} \subseteq \mathbf{L}$ is inconsistent; if $A \subseteq \mathbf{L}$ is inconsistent, so are its supersets $B \subseteq \mathbf{L}$; and L3 holds. See Dietrich (2007) for details.

[9] $\mathbf{L}$ is the smallest set such that (i) $a, b, c, \ldots \in \mathbf{L}$ and (ii) if $p, q \in \mathbf{L}$ then $\neg p, (p \wedge q), (p \vee q), (p \rightarrow q), (p \leftrightarrow q) \in \mathbf{L}$. We drop brackets when there is no ambiguity. Entailment ($\models$) is defined standardly.

$b\}$, $\{\neg a, \neg b, a \rightarrow b\}$, and examples of inconsistent ones are $\{a, \neg a\}$, $\{a, a \rightarrow b, \neg b\}$ and $\{a, b, c \leftrightarrow (a \wedge b), \neg c\}$.

We use classical propositional logic in our examples, but our results also hold for other, more expressive logical languages such as the following:

- *predicate logic*, which includes relation symbols and the quantifiers "there exists ..." and "for all ...";
- *modal logic*, which includes the operators "it's necessary that ..." and "it's possible that ...";
- *deontic logic*, which includes the operators "it's permissible that ..." and "it's obligatory that ...";
- *conditional logic*, which allows the expression of counterfactual or subjunctive conditionals.

Many different propositions that might be considered by a multi-member decision-making body (ranging from legislative committees to expert panels) can be formally represented in an appropriate such language. Crucially, a logical language allows us to capture the fact that, in many decision problems, different propositions, such as the reasons for a particular tenure outcome and the resulting outcome itself, are mutually interconnected.

2.2 The agenda

The *agenda* is the set of propositions on which judgments are to be made; it is a non-empty subset $X \subseteq \mathbf{L}$, where X is a union of proposition-negation pairs $\{p, \neg p\}$ (with p not a negated proposition). For simplicity, we assume that double negations cancel each other out, i.e., $\neg\neg p$ stands for p.[10]

Two important examples are *conjunctive* and *disjunctive* agendas in propositional logic. A conjunctive agenda is $X = \{a_1, \ldots, a_k, c, c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)\}^{+neg}$, where $a_1, \ldots, a_k$ are premises ($k \geq 1$), c is a conclusion, and $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$ is the connection rule. We write Y^{+neg} as an abbreviation for $\{p, \neg p : p \in Y\}$. To define a disjunctive agenda, we replace $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$ with $c \leftrightarrow (a_1 \vee \cdots \vee a_k)$. Conjunctive and disjunctive agendas arise in decision problems in which some outcome (c) is to be decided on the basis of some reasons ($a_1, \ldots, a_k$). In the tenure example above, we have a conjunctive agenda with $k = 2$.[11]

[10] Hereafter, when we write $\neg p$ and p is already of the form $\neg q$, we mean q (rather than $\neg\neg q$).

[11] Although we here interpret connection rules $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$ as *material* biimplications, one may prefer to interpret them as *subjunctive* biimplications (in a conditional logic). This changes the logical relations within conjunctive agendas: more judgment sets are consistent, including $\{\neg a_1, \ldots, \neg a_k, \neg c, \neg(c \leftrightarrow (a_1 \wedge \cdots \wedge a_k))\}$. As a result, our impossibility results (Theorems 2-3 and Corollary 2) do not apply to conjunctive agendas

Other examples are agendas involving conditionals (in propositional or conditional logic) such as $X = \{a, b, a \to b\}^{+neg}$. Here proposition a might state some political goal, proposition $a \to b$ might state what the pursuit of a requires, and proposition b might state the consequence to be drawn. Alternatively, proposition a might be an empirical premise, $a \to b$ a causal hypothesis, and b the resulting prediction.

Finally, we can also represent standard preference aggregation problems within our model. Here we use a predicate logic with a set of constants K representing options ($|K| \geq 3$) and a two-place predicate R representing preferences, where, for any $x, y \in K$, the proposition xRy is interpreted as "x is preferable to y". Now the *preference agenda* is the set $X = \{xRy : x, y \in K\}^{+neg}$ (Dietrich and List 2007a).[12]

The nature of a judgment aggregation problem depends on what propositions are contained in the agenda and how they are interconnected. Our main characterization theorem holds for any agenda of propositions. Our main impossibility theorem holds for a large class of agendas, defined below. We also discuss applications to the important cases of conjunctive and disjunctive agendas.

2.3 Individual and collective judgments

Each individual i's *judgment set* is a subset $A_i \subseteq X$, where $p \in A_i$ means that individual i accepts proposition p . As the agenda typically contains both atomic propositions and compound ones, our definition of a judgment set captures the fact that an individual makes judgments both on free-standing atomic propositions and on their interconnections; and different individuals may disagree with each other on both kinds of propositions.

A judgment set A_i is *consistent* if it is a consistent set of propositions as defined for the logic; A_i is *complete* if it contains a member of each proposition-negation pair $p, \neg p \in X$. A *profile (of individual judgment sets)* is an n-tuple $(A_1, \ldots, A_n)$.

A *(judgment) aggregation rule* is a function F that assigns to each admissible profile $(A_1, \ldots, A_n)$ a collective judgment set $F(A_1, \ldots, A_n) = A \subseteq X$, where $p \in A$ means that the group accepts proposition p. The set of admissible profiles is called the *domain* of F, denoted *Domain* (F). Several results below require the following.

in the revised sense; instead, we obtain stronger possibility results. Analogous remarks hold for disjunctive agendas. See Dietrich (forthcoming).

[12] The entailment relation $\models$ in this logical language is defined by $A \models p$ if and only if $A \cup Z$ entails p in the standard sense of predicate logic, where Z is the set of rationality conditions on preferences $\{(\forall v)vRv, (\forall v_1)(\forall v_2)(\forall v_3)((v_1Rv_2 \wedge v_2Rv_3) \to v_1Rv_3), (\forall v_1)(\forall v_2)(\neg v_1 = v_2 \to (v_1Rv_2 \vee v_2Rv_1))\}$.

Universal Domain. *Domain*(F) is the set of all possible profiles of consistent and complete individual judgment sets.

2.4 Examples of aggregation rules

We give four important examples of aggregation rules satisfying universal domain, as just introduced. The first two rules are defined for any agenda, the last two only for conjunctive (or disjunctive) agendas (the present definitions are simplified, but a generalization is possible).

Propositionwise majority voting. For each $(A_1, \ldots, A_n) \in Domain(F)$, $F(A_1, \ldots, A_n)$ is the set of all propositions $p \in X$ such that more individuals i have $p \in A_i$ than $p \notin A_i$.

Dictatorship of individual i. For each $(A_1, \ldots, A_n) \in Domain(F)$, $F(A_1, \ldots, A_n) = A_i$.

Premise-based voting. For each $(A_1, \ldots, A_n) \in Domain(F)$, $F(A_1, \ldots, A_n)$ is the set containing

- any premise a_j if and only if more i have $a_j \in A_i$ than $a_j \notin A_i$,
- the connection rule $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$,
- the conclusion c if and only if $a_j \in F(A_1, \ldots, A_n)$ for all premises a_j,
- any negated proposition $\neg p$ if and only if $p \notin F(A_1, \ldots, A_n)$.[13]

Here votes are taken only on each premise, and the conclusion is decided by using an exogenously given connection rule.

Conclusion-based voting. For each $(A_1, \ldots, A_n) \in Domain(F)$, $F(A_1, \ldots, A_n)$ is the set containing

- only the conclusion c if more i have $c \in A_i$ than $c \notin A_i$,
- only the negation of the conclusion $\neg c$ otherwise.

Here a vote is taken only on the conclusion, and no collective judgments are made on other propositions.

Dictatorships and premise-based voting always generate consistent and complete collective judgments; propositionwise majority voting sometimes generates inconsistent ones (recall Table 1), and conclusion-based voting always generates incomplete ones (no judgments on the premises).

In debates on the discursive paradox and democratic theory, several arguments have been offered for the superiority of premise-based voting over conclusion-based voting. One such argument draws on a deliberative conception of democracy, which emphasizes that collective decisions on

[13] For a disjunctive agenda, replace "$c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$" with "$c \leftrightarrow (a_1 \vee \cdots \vee a_k)$" and "for all premises a_j" with "for some premise a_j".

conclusions should follow from collectively decided premises (Pettit 2001; Chapman 2002). A second argument draws on the Condorcet jury theorem. If all the propositions are factually true or false and each individual has a probability greater than 1/2 of judging each premise correctly, then, under certain probabilistic independence assumptions, premise-based voting has a higher probability of producing a correct collective judgment on the conclusion than conclusion-based voting (Grofman 1985; Bovens and Rabinowicz 2006; List 2005, 2006). Here we show that, with regard to strategic manipulability, premise-based voting performs worse than conclusion-based voting.

3. NON-MANIPULABILITY

When can an aggregation rule be manipulated by strategic voting? We first introduce a new condition of non-manipulability, not yet game-theoretic. Below we prove that, under certain motivational assumptions about the individuals, our non-manipulability condition is equivalent to a game-theoretic strategy-proofness condition. We also notice that non-manipulability and strategy-proofness may sometimes come apart.

3.1 An example

To give a simple example, we use the language of *incentives* to manipulate, although our subsequent formal analysis focuses on underlying *opportunities* for manipulation; we return to incentives formally in Section 4. Recall the profile in Table 1. Suppose, for the moment, that the three committee members each care only about reaching a collective judgment on the conclusion (c) that agrees with their own individual judgments on the conclusion, and that they do not care about the collective judgments on the premises. What matters to them is the final tenure decision, not the underlying reasons; they are "outcome-oriented", as defined precisely later.

Suppose first the committee uses conclusion-based voting; a vote is taken only on c. Then, clearly, no committee member has an incentive to express an untruthful judgment on c. Individual 1, who wants the committee to accept c, has no incentive to vote against c. Individuals 2 and 3, who want the committee to reject c, have no incentive to vote in favour of c.

But suppose now the committee uses premise-based voting; votes are taken on a and b. What are the members' incentives? Individual 1, who wants the committee to accept c, has no incentive to vote against a or b. But at least one of individuals 2 or 3 has an incentive to vote untruthfully. Specifically, if individuals 1 and 2 vote truthfully, then individual 3 has an incentive to vote untruthfully; and if individuals 1 and 3 vote truthfully, then individual 2 has such an incentive.

To illustrate, assume that individual 2 votes truthfully for a and against b. Then the committee accepts a, regardless of individual 3's vote. So, if individual 3 votes truthfully for b, then the committee accepts b and hence c. But if she votes untruthfully against b, then the committee rejects b and hence c. As individual 3 wants the committee to reject c, she has an incentive to vote untruthfully on b. (In summary, if individual judgments are as in Table 1, voting untruthfully against both a and b weakly dominates voting truthfully for individuals 2 and 3.) Ferejohn (2003) has made this observation informally.

3.2 A non-manipulability condition

To formalize these observations, some definitions are needed. We say that one judgment set, A, *agrees* with another, A^*, on a proposition $p \in X$ if either both or none of A and A^* contains p; A *disagrees* with A^* on p otherwise. Two profiles are *i-variants* of each other if they coincide for all individuals except possibly i.

An aggregation rule F is *manipulable* at the profile $(A_1, \ldots, A_n) \in Domain(F)$ by individual i on proposition $p \in X$ if A_i disagrees with $F(A_1, \ldots, A_n)$ on p, but A_i agrees with $F(A_1, \ldots, A_i^*, \ldots, A_n)$ on p for some i-variant $F(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$.

For example, at the profile in Table 1, premise-based voting is manipulable by individual 3 on c (by submitting $A_3^* = \{\neg a, \neg b, c \leftrightarrow (a \wedge b), \neg c\}$ instead of $A_3 = \{\neg a, b, c \leftrightarrow (a \wedge b), \neg c\}$) and also by individual 2 on c (by submitting $A_2^* = \{\neg a, \neg b, c \leftrightarrow (a \wedge b), \neg c\}$ instead of $A_2 = \{a, \neg b, c \leftrightarrow (a \wedge b), \neg c\}$).

Manipulability thus defined is the existence of an *opportunity* for some individual(s) to manipulate the collective judgment(s) on some proposition(s) by expressing untruthful individual judgments (perhaps on other propositions). The question of when such *opportunities* for manipulation translate into *incentives* for manipulation is a separate question. Whether a rational individual will act on a particular opportunity for manipulation depends on the individual's precise motivation and particularly on how much he or she cares about the various propositions involved in a possible act of manipulation. To illustrate, in our example above, we have assumed that individuals care only about the final tenure decision, implying that they do indeed have incentives to act on their opportunities for manipulation. We discuss this issue in detail when we introduce preferences over judgment sets below.

Our definition of manipulability leads to a corresponding definition of non-manipulability. Let $Y \subseteq X$.

Non-manipulability on Y. F is not manipulable at any profile by any individual on any proposition in Y. Equivalently, for any individual i,

profile $(A_1, \ldots, A_n) \in Domain(F)$ and proposition $p \in Y$, if A_i disagrees with $F(A_1, \ldots, A_n)$ on p, then A_i still disagrees with $F(A_1, \ldots, A_i^*, \ldots, A_n)$ on p for every i-variant $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$.

This definition specifies a family of non-manipulability conditions, one for each $Y \subseteq X$. Non-manipulability on Y requires the absence of opportunities for manipulation on the subset Y of the agenda. If $Y_1 \subseteq Y_2$, then non-manipulability on Y_2 implies non-manipulability on Y_1. If we refer just to "non-manipulability", without adding "on Y", then we mean the default case $Y = X$.

3.3 A characterization result

When is a judgment aggregation rule non-manipulable? We now characterize the class of non-manipulable aggregation rules in terms of an independence condition and a monotonicity condition. Let $Y \subseteq X$.

Independence on Y. For any proposition $p \in Y$ and profiles $(A_1, \ldots, A_n)$, $(A_1^*, \ldots, A_n^*) \in Domain(F)$, if [for all individuals i, $p \in A_i$ if and only if $p \in A_i^*$] then [$p \in F(A_1, \ldots, A_n)$ if and only if $p \in F(A_1^*, \ldots, A_n^*)$].

Monotonicity on Y. For any proposition $p \in Y$, individual i and pair of i-variants $(A_1, \ldots, A_n)$, $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$ with $p \notin A_i$ and $p \in A_i^*$, [$p \in F(A_1, \ldots, A_n)$ implies $p \in F(A_1, \ldots, A_i^*, \ldots, A_n)$].

Weak Monotonicity on Y. For any proposition $p \in Y$, individual i and judgment sets $A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n$, if there exists a pair of i-variants $(A_1, \ldots, A_n)$, $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$ with $p \notin A_i$ and $p \in A_i^*$, then for some such pair [$p \in F(A_1, \ldots, A_n)$ implies $p \in F(A_1, \ldots, A_i^*, \ldots, A_n)$].

Informally, independence on Y states that the collective judgment on each proposition in Y depends only on individual judgments *on that proposition* and not on individual judgments on *other propositions*. Monotonicity (respectively, weak monotonicity) on Y states that an additional individual's support for some proposition in Y never (respectively, not always) reverses the collective acceptance of that proposition (other individuals' judgments remaining fixed).

Again, we have defined families of conditions. If we refer just to "independence" or "(weak) monotonicity", without adding "on Y", then we mean the default case $Y = X$.

THEOREM 1. *Let X be any agenda. For each $Y \subseteq X$, if F satisfies universal domain, the following conditions are equivalent:*

(i) *F is non-manipulable on Y;*
(ii) *F is independent on Y and monotonic on Y;*
(iii) *F is independent on Y and weakly monotonic on Y.*

Without a domain assumption (e.g., for a subdomain of the universal domain), (ii) *and* (iii) *are equivalent, and each implies* (i).[14]

No assumption on the consistency or completeness of collective judgments is needed. The result can be seen as a preference-free analogue in judgment aggregation of a classic characterization of strategy-proof preference aggregation rules by Barberà *et al.* (1993).

In the case of a conjunctive (or disjunctive) agenda, conclusion-based voting is independent and monotonic, hence non-manipulable; premise-based voting is not independent, hence manipulable. But on the set of premises $Y = \{a_1, \ldots, a_k\}^{+neg}$ premise-based voting *is* independent and monotonic (as premise-based voting on those premises is simply equivalent to propositionwise majority voting), and hence it is non-manipulable on Y.

3.4 An impossibility result

Ideally, we want to achieve non-manipulability *simpliciter* and not just on some subset of the agenda. Conclusion-based voting is non-manipulable in this strong sense, but generates incomplete collective judgments. Are there any non-manipulable aggregation rules that generate consistent and complete collective judgments? We now show that, for a general class of agendas, including the agenda in the tenure example above, all non-manipulable aggregation rules satisfying some mild conditions are dictatorial.

To define this class of agendas, we define the notion of *path-connectedness*, a variant of the notion of *total-blockedness* introduced by Nehring and Puppe (2002) (originally in the model of "property spaces").[15] Informally, an agenda of propositions under consideration is *path-connected* if any two propositions in the agenda are logically connected with each other, either directly or indirectly, via a sequence of (conditional) logical entailments.

Formally, proposition p *conditionally entails* proposition q if $\{p, \neg q\} \cup Y$ is inconsistent for some $Y \subseteq X$ consistent with p and with $\neg q$. An agenda X is *path-connected* if, for all contingent[16] propositions $p, q \in X$, there is a sequence $p_1, p_2, \ldots, p_k \in X$ (of length $k \geq 1$) with $p = p_1$ and $q = p_k$ such that p_1 conditionally entails p_2, p_2 conditionally entails $p_3, \ldots, p_{k-1}$ conditionally entails p_k. The class of path-connected agendas includes

[14] Under universal domain, (i), (ii) and (iii) are also equivalent to the conjunction of independence on Y and judgment-set-wise monotonicity on Y, which requires that, for all individuals i and all i-variants $(A_1, \ldots, A_n)$, $(A_1^*, \ldots, A_i^*, \ldots, A_n^*) \in Domain(F)$, if $A_i^* = F(A_1, \ldots, A_n)$ then $F(A_1^*, \ldots, A_i^*, \ldots, A_n^*) \cap Y = F(A_1, \ldots, A_n) \cap Y$.

[15] For a compact logic, *path-connectedness* is equivalent to total blockedness; in the general case, path-connectedness is weaker.

[16] We call a proposition $p \in \mathbf{L}$ *contingent* if both $\{p\}$ and $\{\neg p\}$ are consistent.

conjunctive and disjunctive agendas (see the Appendix) and the preference agenda (Nehring 2003; Dietrich and List 2007a), which can be used to represent Condorcet–Arrow preference aggregation problems.

Consider the following conditions on an aggregation rule in addition to universal domain.

Collective Rationality. For any profile $(A_1, \ldots, A_n) \in Domain(F)$, $F(A_1, \ldots, A_n)$ is consistent and complete.[17]

Responsiveness. For any contingent proposition $p \in X$, there exist two profiles $(A_1, \ldots, A_n)$, $(A_1^*, \ldots, A_n^*) \in Domain(F)$ such that $p \in F(A_1, \ldots, A_n)$ and $p \notin F(A_1^*, \ldots, A_n^*)$.

THEOREM 2. *For a path-connected agenda X (e.g., a conjunctive, disjunctive or preference agenda), an aggregation rule F satisfies universal domain, collective rationality, responsiveness and non-manipulability if and only if F is a dictatorship of some individual.*

For the important case of compact logical languages, this result also follows from Theorem 1 above and Nehring and Puppe's (2002) characterization of monotonic and independent aggregation rules for totally blocked agendas.[18] Theorem 2 is the judgment aggregation analogue of the Gibbard–Satterthwaite theorem on preference aggregation, which shows that dictatorships are the only strategy-proof social choice functions that satisfy universal domain, have three or more options in their range and always produce a determinate winner (Gibbard 1973; Satterthwaite 1975). Below we restate Theorem 2 using a game-theoretic strategy-proofness condition.

In the special case of the preference agenda, however, there is an interesting disanalogy between Theorem 2 and the Gibbard–Satterthwaite theorem. As a collectively rational judgment aggregation rule for the preference agenda represents an Arrowian social welfare function, Theorem 2 establishes an impossibility result on the non-manipulability of social welfare functions (generating orderings as in Arrow's framework) as opposed to social choice functions (generating winning options as in the Gibbard–Satterthwaite framework); for a related result, see Bossert and Storcken (1992).

If the agenda is not path-connected, then there may exist non-dictatorial aggregation rules satisfying all of Theorem 2's conditions;

[17] Although completeness is conventionally called a rationality requirement, one may consider consistency more important. But if the agenda includes all those propositions on which collective judgments are (practically) required, completeness seems reasonable. Below we discuss relaxing it.

[18] Nehring and Puppe's result implies that the theorem's agenda assumption is maximally weak.

examples of such agendas are not only trivial agendas (containing a single proposition-negation pair or several logically independent such pairs), but also agendas involving conditionals, including the simple example $X = \{a, b, a \rightarrow b\}^{+neg}$ (Dietrich forthcoming).

By contrast, for *atomically closed* or *atomic* agendas, special cases of path-connected agendas with very rich logical connections, an even stronger impossibility result holds, in which Theorem 2's responsiveness condition is significantly weakened.[19]

Weak Responsiveness. The aggregation rule is non-constant. Equivalently, there exist two profiles $(A_1, \ldots, A_n), (A_1^*, \ldots, A_n^*) \in Domain(F)$ such that $F(A_1, \ldots, A_n) \neq F(A_1^*, \ldots, A_n^*)$.

THEOREM 3. *For an atomically closed or atomic agenda X, an aggregation rule F satisfies universal domain, collective rationality, weak responsiveness and non-manipulability if and only if F is a dictatorship of some individual.*

Given Theorem 1 above, this result follows immediately from theorems by Pauly and van Hees (2006) (for atomically closed agendas) and Dietrich (2006) (for atomic ones).

3.5 Avoiding the impossibility result

To find non-manipulable and non-dictatorial aggregation rules, we must relax at least one condition in Theorems 2 or 3. Non-responsive rules are usually unattractive. Permitting inconsistent collective judgments also seems unattractive. But the following may sometimes be defensible.

Incompleteness. For a conjunctive or disjunctive agenda, conclusion-based voting is non-manipulable. It generates incomplete collective judgments and is only weakly responsive; this may be acceptable when no collective judgments on the premises are required. More generally, *propositionwise supermajority rules* – requiring a supermajority of a particular size (or even unanimity) for the acceptance of a proposition – are consistent and non-manipulable (by Theorem 1), again at the expense of violating completeness as neither member of a pair $p, \neg p \in X$ might obtain the required supermajority. For a finite agenda (or compact logical languages), a supermajority rule requiring at least m votes for the acceptance of any proposition guarantees collective consistency if and only

[19] Agenda X is *atomically closed* if (i) X belongs to classical propositional logic, (ii) if an atomic proposition a occurs in some $p \in X$ then $a \in X$, and (iii) for any atomic propositions $a, b \in X$, we have $a \wedge b, a \wedge \neg b, \neg a \wedge b, \neg a \wedge \neg b \in X$ (Pauly and van Hees 2006). X is *atomic* if $\{\neg p : p \text{ is an atom of } X\}$ is inconsistent, where $p \in X$ is an *atom* of X if p is consistent but inconsistent with some member of each pair $q, \neg q \in X$ (Dietrich 2006). In Theorem 3, X must contain two (or more) contingent propositions p and q, with p not equivalent to q or $\neg q$.

if $m > n - n/z$, where z is the size of the largest minimal inconsistent set $Z \subseteq X$ (Dietrich and List 2007b; List 2004).

Domain restriction. By suitably restricting the domain of propositionwise majority voting, this rule becomes consistent; it is also non-manipulable as it is independent and monotonic. This result holds, for example, for the domain of all profiles of consistent and complete individual judgment sets satisfying the structure condition of *unidimensional alignment* (List 2003).[20] Informally, unidimensional alignment requires that the individuals can be aligned from left to right (under any interpretation of "left" and "right") such that, for each proposition on the agenda, the individuals accepting the proposition are either exclusively to the left, or exclusively to the right, of those rejecting it. This structure condition captures a shared unidimensional conceptualization of the decision problem by the decision-makers. In debates on deliberative democracy, it is sometimes hypothesized that group deliberation may reduce disagreement so as to bring about such a shared unidimensional conceptualization (Miller 1992; Dryzek and List 2003), sometimes also described as a "meta-consensus" (List 2002a).

4. STRATEGY-PROOFNESS

Non-manipulability is not yet a game-theoretic concept. We now define strategy-proofness, a game-theoretic concept that depends on individual preferences (over judgment sets held by the group). We identify assumptions on individual preferences that render strategy-proofness equivalent to non-manipulability and discuss the plausibility of these assumptions.

4.1 Preference relations over judgment sets

We interpret a judgment aggregation problem as a game with n players (the individuals).[21] The game form is given by the aggregation rule: each individual's possible actions are the different judgment sets the individual can submit to the aggregation rule (which may or may not coincide with the individual's true judgment set); the outcomes are the collective judgment sets generated by the aggregation rule.

To specify the game fully, we assume that each individual, in addition to holding a true judgment set A_i, also has a preference relation $\succsim_i$ over all possible outcomes of the game, i.e., over all possible collective judgment sets of the form $A \subseteq X$. For any two judgment sets, $A, B \subseteq X$, $A \succsim_i B$

[20] For a related result on preference aggregation, see Saporiti and Thomé (2005).

[21] For an earlier version of this game-theoretic interpretation of judgment aggregation, the notion of closeness-respecting preferences over judgment sets, and a sufficient condition for strategy-proofness (in a sequential context), see List (2002b, 2004).

means that individual i weakly prefers the group to endorse A as the collective judgment set rather than B. We assume that $\succsim_i$ is reflexive and transitive, but do not require it to be complete. [22] Individuals need not be able to rank all pairs of judgment sets relative to each other; in principle, our model allows studying a further relaxation of these conditions.

What preferences over collective judgment sets can we expect an individual i to hold when i's judgment set is A_i? The answer is not straightforward, and it may even be difficult to say *anything* about i's preferences on the basis of A_i alone. To illustrate this, consider first a single proposition p, say, "CO_2 emissions lead to global warming". If individual i judges that p (i.e., $p \in A_i$), it does not necessarily follow that i wants the group to judge that p. Just imagine that i owns an oil company which benefits from low taxes on CO_2 emissions, and that taxes are increased if and only if the group judges that p. In general, accepting p and wanting the group to accept p are conceptually distinct (though the literature is often unclear about this distinction). Whether acceptance and desire of group acceptance happen to coincide in a particular case is an empirical question.[23] There are important situations in which the two may indeed be reasonably expected to coincide. An important example is that of *epistemically motivated* individuals: here each individual prefers group judgments that she considers closer to the truth, where she may consider her own judgments as the truth. A *non-epistemically motivated* individual prefers judgment sets for reasons other than the truth, for example because she personally benefits from group actions resulting from the collective endorsement of some judgment sets rather than others.[24]

We now give examples of possible assumptions (empirical claims) on how the individuals' preferences are related to their judgment sets. Which of these assumptions is correct depends on the group of individuals and the aggregation problem in question. Different assumptions capture

[22] $\succsim_i$ is: *reflexive* if, for any A, $A \succsim_i A$; *transitive* if, for any A, B, C, $A \succsim_i B$ and $B \succsim_i C$ implies $A \succsim_i C$; *complete* if, for any distinct A, B, $A \succsim_i B$ or $B \succsim_i A$.

[23] This argument identifies accepting with believing, thus interpreting judgment sets as (binary) belief sets, and judgment aggregation as the aggregation of (binary) belief sets into group belief sets. Although this interpretation is standard, other interpretations are possible. If accepting means desiring, judgment aggregation is the aggregation of (binary) desire sets into group desire sets. It is then more plausible that i wants the group to accept (desire) the propositions that i accepts (desires).

[24] Even non-epistemically motivated individuals may sometimes prefer group judgments that match their own individual judgments. Suppose each individual is motivated by her desires over outcomes of group actions, which depend on the state of the world. Suppose, further, all individuals hold the same desires over outcomes but different beliefs about the state of the world, and each individual is convinced that her own beliefs are true and that their collective acceptance would lead to the desired outcomes. Such individuals may want the group judgments to match their individual judgments, but mainly to satisfy their desires over outcomes rather than to bring about true group beliefs.

different motivations of the individuals, as illustrated above. Specifically, the assumption of "unrestricted" preferences captures the case where an individual's preferences are not in any systematic way linked to her judgments; the assumption of "top-respecting" preferences and the stronger one of "closeness-respecting" preferences capture situations in which agents would like group judgments to agree with their own judgments. We use a function C that assigns to each possible judgment set A_i a non-empty set $C(A_i)$ of (reflexive and transitive) preference relations that are considered "compatible" with A_i (i.e., possible given A_i). Our examples of preference assumptions can be stated formally as follows (in increasing order of strength).

Unrestricted preferences. For each A_i, $C(A_i)$ is the set of all preference relations $\succsim$ (regardless of A_i).

Top-respecting preferences. For each A_i, $C(A_i)$ is the set of all preference relations $\succsim$ for which A_i is a most preferred judgment set, i.e., $C(A_i) = \{\succsim: A_i \succsim B \text{ for all judgment sets } B\}$.

To define "closeness-respecting" preferences, we say that a judgment set B is *at least as close* to A_i on some $Y \subseteq X$ as another judgment set B^* if, for all propositions $p \in Y$, if B^* agrees with A_i on p, then B also agrees with A_i on p. For example, $\{\neg a, b, c \leftrightarrow (a \wedge b), \neg c\}$ is at least as close to $\{a, b, c \leftrightarrow (a \wedge b), c\}$ on X as $\{\neg a, \neg b, c \leftrightarrow (a \wedge b), \neg c\}$,[25] whereas $\{\neg a, b, c \leftrightarrow (a \wedge b), \neg c\}$ and $\{a, \neg b, c \leftrightarrow (a \wedge b), \neg c\}$ are unranked in terms of relative closeness to $\{a, b, c \leftrightarrow (a \wedge b), c\}$ on X. We say that a preference relation $\succsim$ respects closeness to A_i on Y if, for any two judgment sets B and B^*, if B is at least as close to A_i as B^* on Y, then $B \succsim B^*$.

Closeness-respecting preferences on Y (for some $Y \subseteq X$). For each A_i, $C(A_i)$ is the set of all preference relations $\succsim$ that respect closeness to A_i on Y, and we write $C = C_Y$.

In the important case $Y = X$, we drop the reference "on Y" and speak of closeness-respecting preferences *simpliciter*. One element of $C_X(A_i)$ is the (complete) preference relation induced by the Hamming distance to A_i.[26] Below we analyse the important cases of "reason-oriented" and

[25] In fact, it is "closer", where "closer than" is the strong component of "at least as close as".

[26] The Hamming distance between two judgment sets B and B^* is $d(B, B^*) := |\{p \in X : B \text{ and } B^* \text{ disagree on } p\}|$. The preference relation $\succeq$ induced by Hamming distance to A_i is defined, for any B, B^*, by $[B \succeq B^*$ if and only if $d(B, A_i) \leq d(B^*, A_i)]$. For the preference agenda, a preference relation $\succeq$ over judgment sets (each representing a preference ordering over the option set K) represents a meta-preference over preference orderings. Bossert and Storcken (1992) use the Kemeny distance between preference orderings to obtain such a meta-preference. For related work on distances between preferences and theories, see Baigent (1987) and Schulte (2005), respectively.

"outcome-oriented" preferences, where Y is given by particular subsets of X. Generally, if $Y_1 \subseteq Y_2$, then, for all A_i, $C_{Y_1}(A_i) \subseteq C_{Y_2}(A_i)$.

4.2 A strategy-proofness condition

Given a specification of the function C, an aggregation rule is strategy-proof for C if, for any profile, any individual and any preference relation compatible with the individual's judgment set (according to C), the individual (weakly) prefers the outcome of expressing her judgment set truthfully to any outcome that would result from misrepresenting her judgment set.

Strategy-proofness for C. For any individual i, profile $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ and preference relation $\succsim_i \in C(A_i)$, $F(A_1, \ldots, A_n) \succsim_i F(A_1, \ldots, A_i^*, \ldots, A_n)$ for every i-variant $(A_1, \ldots, A_i^*, \ldots, A_n) \in \mathit{Domain}(F)$.[27]

If the aggregation rule F has the universal domain, then strategy-proofness implies that truthfulness is a weakly dominant strategy for every individual.[28] Our definition of strategy-proofness (generalizing List 2002b, 2004) is similar to Gibbard's (1973) and Satterthwaite's (1975) classical one and related to other definitions of strategy-proofness in the literature on preference aggregation (particularly, for C_X, those by Barberà *et al.* (1993, 1997) and Nehring and Puppe (2002), employing the notion of generalized single-peaked preferences).

As in the case of non-manipulability above, we have defined a family of strategy-proofness conditions, one for each specification of C. This means that different motivational assumptions about the individuals lead to different strategy-proofness conditions. If individuals have very restrictive preferences over possible judgment sets, then strategy-proofness is easier to achieve than if their preferences are largely unrestricted. Formally, if two functions C_1 and C_2 are such that $C_1 \subseteq C_2$ (i.e., for each A_i, $C_1(A_i) \subseteq C_2(A_i)$), then strategy-proofness for C_1 is less demanding than (i.e., implied by) strategy-proofness for C_2. The more preference relations are compatible with each individual judgment set, the more demanding is the corresponding requirement of strategy-proofness.

[27] Our definition of strategy-proofness can be generalized by admitting a different function C_i for each individual i. This removes a homogeneity assumption, whereby, if individuals i and j hold the same judgment set $A_i = A_j$, then their preference relations fall into the same set $C(A_i) = C(A_j)$. The homogeneity assumption is undemanding when $C(A_i)$ is large.

[28] This interpretation of strategy-proofness holds for product domains. For certain subdomains of the universal domain (i.e., non-product domains), we do not have a strictly well-defined game, but our definition of strategy-proofness remains applicable and can be reinterpreted as one of "conditional strategy-proofness" for non-product domains, as discussed by Saporiti and Thomé (2005).

4.3 The equivalence of strategy-proofness and non-manipulability

What is the logical relation between non-manipulability as defined above and strategy-proofness? We show that, if preferences are closeness-respecting (on some $Y \subseteq X$), then an equivalence between these two concepts arises. Let X be any agenda.

THEOREM 4. *For each $Y \subseteq X$, F is strategy-proof for C_Y if and only if F is non-manipulable on Y.*

In other words, for any subset Y of the agenda X (including the case $Y = X$), strategy-proofness of an aggregation rule for closeness-respecting preferences on Y is equivalent to non-manipulability on the propositions in Y. In particular, strategy-proofness for closeness-respecting preferences *simpliciter* is equivalent to non-manipulability *simpliciter*. This also implies that, for unrestricted or top-respecting preferences, strategy-proofness is more demanding than our default condition of non-manipulability, whereas, for closeness-respecting preferences on some $Y \subseteq X$, it is less demanding.

Given the equivalence result of Theorem 4, we can now state corollaries of Theorems 1 and 2 above for strategy-proofness:[29]

COROLLARY 1. *For each $Y \subseteq X$, if F satisfies universal domain, the following conditions are equivalent:*

(i) *F is strategy-proof for C_Y;*
(ii) *F is independent on Y and monotonic on Y;*
(iii) *F is independent on Y and weakly monotonic on Y.*

Without a domain assumption (e.g., for a subdomain of the universal domain), (ii) *and* (iii) *are equivalent, and each implies* (i).

COROLLARY 2. *For a path-connected agenda X (e.g., a conjunctive, disjunctive or preference agenda), an aggregation rule F satisfies universal domain, collective rationality, responsiveness and strategy-proofness for C_X if and only if F is a dictatorship of some individual.*

Corollary 2 is a judgment aggregation analogue of Nehring and Puppe's (2002) characterization of strategy-proof social choice functions in the model of "property spaces".[30] The negative part of corollary 2 (i.e., if an aggregation rule satisfies the conditions, then it is a dictatorship)

[29] Our remarks on Theorems 1 and 2 above also apply to Corollaries 1 and 2.

[30] For compact logics, it follows from their result via Corollary 1. As noted, a disanalogy lies in the aggregation rule's different informational input. In Barberà *et al.* (1993, 1997) and Nehring and Puppe (2002), each individual submits a preference relation, here a single judgment set. Under some conditions, judgment sets can be associated with peaks of preference relations.

holds not only for closeness-respecting preferences (C_X) but for any preference specification C at least as broad as C_X, i.e., $C_X \subseteq C$, as strategy-proofness for C then implies strategy-proofness for C_X. The positive part of corollary 2 (i.e., if an aggregation rule is a dictatorship, then it satisfies the conditions) holds for any preference specification C allowing only top-respecting preferences, i.e., for any C such that, if $\succsim \in C(A_i)$, then $A_i \succsim B$ for all judgment sets B; otherwise a dictatorship, although non-manipulable, is not strategy-proof (to see this point, recall the example of the oil company in Section 4.1).

In summary, if the individuals' preferences over judgment sets are unrestricted, top-respecting or closeness-respecting, we obtain a negative result. Moreover, in analogy with Theorem 3 above, for atomically closed or atomic agendas, we get an impossibility result even if we weaken responsiveness to the requirement of a non-constant aggregation rule.

5. OUTCOME- AND REASON-ORIENTED PREFERENCES

As we have introduced families of strategy-proofness and non-manipulability conditions, it is interesting to consider some less demanding conditions within these families. If we demand strategy-proofness for $C = C_X$, equivalent to non-manipulability *simpliciter*, this precludes all incentives for manipulation, where individuals have closeness-respecting preferences. But individual preferences may sometimes fall into a more restricted set: they may be closeness-respecting on some subset $Y \subseteq X$, in which case it is sufficient to require strategy-proofness for C_Y. As an illustration, we now apply these ideas to the case of a conjunctive (analogously disjunctive) agenda.

5.1 Definition

Let X be a conjunctive (or disjunctive) agenda. Two important cases of closeness-respecting preferences on Y are the following.

Outcome-oriented preferences. $C = C_{Y_{outcome}}$, where $Y_{outcome} = \{c\}^{+neg}$.

Reason-oriented preferences. $C = C_{Y_{reason}}$, where $Y_{reason} = \{a_1, \ldots, a_k\}^{+neg}$.

An individual with outcome-oriented preferences cares only about achieving a collective judgment on the conclusion that matches her own judgment, regardless of the premises. Such preferences make sense if only the conclusion but not the premises have consequences the individual cares about. An individual with reason-oriented preferences cares only about achieving collective judgments on the premises that match her own judgments, regardless of the conclusion. Such preferences make sense if the individual gives primary importance to the reasons given in support of outcomes, rather than the outcomes themselves, or if the group's

judgments on the premises have important consequences themselves that the individual cares about (such as setting precedents for future decisions). Proponents of a deliberative conception of democracy often argue that the motivational assumption of reason-oriented preferences is appropriate in deliberative settings (for a discussion, see Elster 1986; Goodin 1986). Economists, by contrast, assume that in many settings outcome-oriented preferences are the more accurate motivational assumption. Ultimately, it is an empirical question what preferences are triggered by various settings.

To illustrate, consider premise-based voting and the profile in Table 1. Individual 3's judgment set is $A_3 = \{\neg a, b, \neg c, r\}$, where $r = c \leftrightarrow (a \wedge b)$. If all individuals are truthful, the collective judgment set is $A = \{a, b, c, r\}$. If individual 3 untruthfully submits $A_3^* = \{\neg a, \neg b, \neg c, r\}$ and individuals 1 and 2 are truthful, the collective judgment set is $A^* = \{a, \neg b, \neg c, r\}$. Now A^* is closer to A_3 than A on $Y_{outcome} = \{c\}^{+neg}$, whereas A is closer to A_3 than A^* on $Y_{reason} = \{a, b\}^{+neg}$. So, under outcome-oriented preferences, individual 3 (at least weakly) prefers A^* to A, whereas, under reason-oriented preferences, individual 3 (at least weakly) prefers A to A^*.

5.2 The strategy-proofness of premise-based voting for reason-oriented preferences

As shown above, conclusion-based voting is strategy-proof for C_X and hence also for $C_{Y_{reason}}$ and $C_{Y_{outcome}}$. Premise-based voting is not strategy-proof for C_X and neither for $C_{Y_{outcome}}$, as can easily be seen from our first example of manipulation. But the following holds.

Proposition 1. *For a conjunctive or disjunctive agenda X, premise-based voting is strategy-proof for* $C_{Y_{reason}}$.

This result is interesting from a deliberative democracy perspective. If individuals have reason-oriented preferences in deliberative settings, as sometimes argued by proponents of a deliberative conception of democracy, then premise-based voting is strategy-proof in such settings. But if individuals have outcome-oriented preferences, then the aggregation rule advocated by deliberative democrats is vulnerable to strategic manipulation, posing a challenge to the deliberative democrats' view that truthfulness can easily be achieved under their preferred aggregation rule.

5.3 The strategic equivalence of premise- and conclusion-based voting for outcome-oriented preferences

Surprisingly, if individuals have outcome-oriented preferences, then premise- and conclusion-based voting are strategically equivalent in the following sense. For any profile, there exists, for each of the two rules, a (weakly) dominant-strategy equilibrium leading to the same collective

judgment on the conclusion. To state this result formally, some definitions are needed.

Under an aggregation rule F, for individual i with preference ordering $\succsim_i$, submitting the judgment set B_i (which may or may not coincide with individual i's true judgment set A_i) is a *weakly dominant strategy* if, for every profile $(B_1, \ldots, B_i, \ldots, B_n) \in Domain(F)$, $F(B_1, \ldots, B_i, \ldots, B_n) \succsim_i F(B_1, \ldots, B_i^*, \ldots, B_n)$ for every i-variant $(B_1, \ldots, B_i^*, \ldots, B_n) \in Domain(F)$.

Two aggregation rules F and G with identical domain are *strategically equivalent* on $Y \subseteq X$ for C if, for every profile $(A_1, \ldots, A_n) \in Domain(F) = Domain(G)$ and preference relations $\succsim_1 \in C(A_1), \ldots, \succsim_n \in C(A_n)$, there exist profiles $(B_1, \ldots, B_n), (C_1, \ldots, C_n) \in Domain(F) = Domain(G)$ such that

(i) for each individual i, submitting B_i is a weakly dominant strategy under rule F and submitting C_i is a weakly dominant strategy under rule G;
(ii) $F(B_1, \ldots, B_n)$ and $G(C_1, \ldots, C_n)$ agree on every proposition $p \in Y$.

THEOREM 5. *For a conjunctive or disjunctive agenda X, premise- and conclusion-based voting are strategically equivalent on $Y_{outcome} = \{c\}^{+neg}$ for $C_{Y_{outcome}}$.*

Despite the differences between premise- and conclusion-based voting, if individuals have outcome-oriented preferences and act on appropriate weakly dominant strategies, the two rules generate identical collective judgments on the conclusion. This is surprising as premise- and conclusion-based voting are regarded in the literature as two diametrically opposed aggregation rules.

6. CONCLUDING REMARKS

As judgment aggregation problems arise in many real-world decision-making bodies, it is important to understand which judgment aggregation rules are vulnerable to manipulation and which not. We have introduced a non-manipulability condition for judgment aggregation and characterized the class of non-manipulable judgment aggregation rules. Non-manipulability rules out the existence of *opportunities* for manipulation by the untruthful expression of individual judgments. We have then defined a game-theoretic strategy-proofness condition and shown that, under some (but not all) motivational assumptions, it is equivalent to non-manipulability, as defined earlier. For these motivational assumptions, our characterization of non-manipulable aggregation rules has allowed us to characterize all strategy-proof aggregation rules. Strategy-proofness rules out the existence of *incentives* for manipulation. Crucially, if individuals do not generally want the group to make collective

judgments that match their own individual judgments, the concepts of non-manipulability and strategy-proofness may come significantly apart.

We have also proved an impossibility result that is the judgment aggregation analogue of the classical Gibbard–Satterthwaite theorem on preference aggregation. For the class of path-connected agendas, including conjunctive, disjunctive and preference agendas, all non-manipulable aggregation rules satisfying some mild conditions are dictatorial. The impossibility result becomes even stronger for agendas with particularly rich logical connections between propositions.

To avoid this impossibility, we have suggested that permitting incomplete collective judgments or domain restrictions are the most promising routes. For example, conclusion-based voting is strategy-proof, but violates completeness. Another way to avoid the impossibility is to relax non-manipulability or strategy-proofness itself. Both conditions fall into more general families of conditions of different strength. Instead of requiring non-manipulability on the entire agenda of propositions, we may require non-manipulability only on some subset of the agenda. Premise-based voting, for example, is non-manipulable on the set of premises, but not non-manipulable *simpliciter*. Whether such a weaker non-manipulability condition is sufficient in practice depends on how worried we are about possible opportunities for manipulation on propositions outside the subset of the agenda for which non-manipulability holds. Likewise, instead of requiring strategy-proofness for a large class of individual preferences over judgment sets, we may require strategy-proofness only for a restricted class of preferences, for example for "outcome-" or "reason-oriented" preferences. Premise-based voting, for example, is strategy-proof for "reason-oriented" preferences. Whether such a weaker strategy-proofness condition is sufficient in practice depends on the motivations of the decision-makers.

Finally, we have shown that, for "outcome-oriented" preferences, premise- and conclusion-based voting are strategically equivalent. They generate the same collective judgment on the conclusion if individuals act on appropriate weakly dominant strategies.

Our results raise questions about a prominent position in the literature, according to which premise-based voting is superior to conclusion-based voting from a deliberative democracy perspective. We have shown that, with respect to non-manipulability and strategy-proofness, conclusion-based voting outperforms premise-based voting. This result could be generalized beyond conjunctive and disjunctive agendas.

Until now, comparisons between judgment aggregation and preference aggregation have focused mainly on Condorcet's paradox and Arrow's theorem. With this paper, we hope to inspire further research on strategic voting and a game-theoretic perspective in a judgment aggregation context. An important challenge is the development of models

of *deliberation* on interconnected propositions – where individuals not only "feed" their judgments into some aggregation rule, but where they deliberate about the propositions prior to making collective judgments – and the study of the strategic aspects of such deliberation. We leave this challenge for further work.

A. APPENDIX

Proof of Theorem 1. Let $Y \subseteq X$. We prove first that (ii) and (iii) are equivalent, then that (ii) implies (i), and then that, given universal domain, (i) implies (ii).

(ii) implies (iii). Trivial as monotonicity on Y implies weak monotonicity on Y.

(iii) implies (ii). Suppose F is independent on Y and weakly monotonic on Y.

To show monotonicity on Y, note that in the requirement defining weak monotonicity on Y one may, by independence on Y, replace "for some such pair" by "for all such pairs". The modified requirement is equivalent to monotonicity on Y.

(ii) implies (i). Suppose F is independent on Y and monotonic on Y. To show non-manipulability on Y, consider any proposition $p \in Y$, individual i, and profile $(A_1, \ldots, A_n) \in Domain(F)$, such that $F(A_1, \ldots, A_n)$ disagrees with A_i on p. Take any i-variant $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$. We have to show that $F(A_1, \ldots, A_i^*, \ldots, A_n)$ still disagrees with A_i on p. Assume first that A_i and A_i^* agree on p. Then in both profiles $(A_1, \ldots, A_n)$ and $(A_1, \ldots, A_i^*, \ldots, A_n)$ exactly the same individuals accept p. Hence, by independence on Y, $F(A_1, \ldots, A_i^*, \ldots, A_n)$ agrees with $F(A_1, \ldots, A_n)$ on p, hence disagrees with A_i on p. Now assume A_i^* disagrees with A_i on p, i.e., agrees with $F(A_1, \ldots, A_n)$ on p. Then, by monotonicity on Y, $F(A_1, \ldots, A_i^*, \ldots, A_n)$ agrees with $F(A_1, \ldots, A_n)$ on p, i.e., disagrees with A_i on p.

(i) implies (ii). Now assume universal domain, and let F be non-manipulable on Y. To show monotonicity on Y, consider any proposition $p \in Y$, individual i, and pair of i-variants $(A_1, \ldots, A_n)$, $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$ with $p \notin A_i$ and $p \in A_i^*$. If $p \in F(A_1, \ldots, A_n)$, then A_i disagrees on p with $F(A_1, \ldots, A_n)$, hence also with $F(A_1, \ldots, A_i^*, \ldots, A_n)$ by non-manipulability on Y. So $p \in F(A_1, \ldots, A_i^*, \ldots, A_n)$. To show independence on Y, consider any proposition $p \in Y$ and profiles $(A_1, \ldots, A_n)$, $(A_1^*, \ldots, A_n^*) \in Domain(F)$ such that, for all individuals i, A_i and A_i^* agree on p. We have to show that $F(A_1, \ldots, A_n)$ and $F(A_1^*, \ldots, A_n^*)$ agree on p. Starting with the profile $(A_1, \ldots, A_n)$, we replace first A_1 by A_1^*, then A_2 by $A_2^*, \ldots$, then A_n by A_n^*. By universal domain, each replacement leads to a profile still in $Domain(F)$. We now show that each replacement preserves the collective judgment about p. Assume for contradiction that for individual i

replacement of A_i by A_i^* changes the collective judgment about p. Since A_i and A_i^* agree on p but the respective outcomes for A_i and for A_i^* disagree on p, either A_i or A_i^* (but not both) disagrees with the respective outcome. This is a contradiction, since it allows individual i to manipulate: in the first case by submitting A_i^* with genuine judgment set A_i, in the second case by submitting A_i with genuine judgment set A_i^*. Since no replacement has changed the collective judgment about p, it follows that $F(A_1, \ldots, A_n)$ and $F(A_1, \ldots, A_n)$ agree on p, which proves independence on Y.

For any propositions p, q, we write $p \models^* q$ to mean that p *conditionally entails* q.

Proof that conjunctive and disjunctive agendas are path-connected. Let X be the conjunctive agenda $X = \{a_1, \neg a_1, \ldots, a_k, \neg a_k, c, \neg c, r, \neg r\}$, where $k \geq 1$ and r is the connection rule $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$. (The proof for a disjunctive agenda is analogous.) We have to show that for any $p, q \in X$ there is a sequence $p = p_1, p_2, \ldots, p_k = q$ in X ($k \geq 1$) such that $p_1 \models^* p_2$, $p_2 \models^* p_3, \ldots, p_{k-1} \models^* p_k$. To show this, it is sufficient to prove that

(1) $p \models^* q$ for any propositions $p, q \in X$ of *different types*,

where a proposition is of type 1 if it is a possibly negated premise $(a_1, \neg a_1, \ldots, a_k, \neg a_k)$, of type 2 if it is the possibly negated conclusion $(c, \neg c)$ and of type 3 if it is the possibly negated connection rule $(r, \neg r)$. The reason is (in short) that, if (1) holds, then, for any $p, q \in X$ of the *same* type, taking any $s \in X$ of a different type, there is by (1) a path connecting p to s and a path connecting s to q; the concatenation of both paths connects p to q, as desired. As $p \models^* q$ if and only if $\neg q \models^* \neg p$ (use both times the same Y), claim (1) is equivalent to

(2) $p \models^* q$ for any propositions $p, q \in X$ such that p has smaller type than q.

We show (2) by going through the different cases (where $j \in \{1, \ldots, k\}$):

From type 2 to type 3: we have $c \models^* r$ and $\neg c \models^* \neg r$ (take $Y = \{a_1 \ldots, a_k\}$ both times), and $c \models^* \neg r$ and $\neg c \models^* r$ (take $Y = \{\neg a_1\}$ both times).

From type 1 to type 2: we have $a_j \models^* c$ and $\neg a_j \models^* \neg c$ (take $Y = \{r, a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_k\}$ both times), and $a_j \models^* \neg c$ and $\neg a_j \models^* c$ (take $Y = \{\neg r, a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_k\}$ both times);

From type 1 to type 3: we have $a_j \models^* r$ and $\neg a_j \models^* \neg r$ (take $Y = \{c, a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_k\}$ both times), and $a_j \models^* \neg r$ and $\neg a_j \models^* r$ (take $Y = \{\neg c, a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_k\}$ both times).

Proof of Theorem 2. Let X be path-connected. If F is dictatorial, it obviously satisfies universal domain, collective rationality, responsiveness

and non-manipulability. Now suppose F has all these properties, hence is also independent and monotonic by Theorem 1. We show that F is dictatorial. If X contains no contingent proposition, F is trivially dictatorial (where each individual is a dictator). From now on, suppose X is not of this degenerate type. For any consistent set $Z \subseteq X$, let A_Z be some consistent and complete judgment set such that $Z \subseteq A_Z$ (which exists by L1–L3).

Claim 1. F satisfies the unanimity principle: for any $p \in X$ and any $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$, if $p \in A_i$ for each i then $p \in F(A_1, \ldots, A_n)$.

Consider any $p \in X$ and $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ such that $p \in A_i$ for every i. Since the sets A_i are consistent, p is consistent. If $\neg p$ is inconsistent (i.e., p is a tautology), $p \in F(A_1, \ldots, A_n)$ by collective rationality. Now suppose $\neg p$ is consistent. As each of $p, \neg p$ is consistent, p is contingent. So, by responsiveness, there exists a profile $(B_1, \ldots, B_n) \in \mathit{Domain}(F)$ such that $p \in F(B_1, \ldots, B_n)$. In $(B_1, \ldots, B_n)$ we now replace one by one each judgment set B_i by A_i, until we obtain the profile $(A_1, \ldots, A_n)$. Each replacement preserves the collective acceptance of p, either by monotonicity (if $p \notin B_i$) or by independence (if $p \in B_i$). So $p \in F(A_1, \ldots, A_n)$, as desired.

Claim 2. F is systematic: there exists a set $\mathcal{W}$ of ("winning") coalitions $C \subseteq N$ such that, for every $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$, $F(A_1, \ldots, A_n) = \{p \in X : \{i : p \in A_i\} \in \mathcal{W}\}$.

For each contingent $p \in X$, let $\mathcal{W}_p$ be the set all subsets $C \subseteq N$ such that $p \in F(A_1, \ldots, A_n)$ for some (hence by independence any) $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ with $\{i : p \in A_i\} = C$. Consider any contingent $p, q \in X$. We prove that $\mathcal{W}_p = \mathcal{W}_q$. Suppose $C \in \mathcal{W}_p$, and let us show that $C \in \mathcal{W}_q$; this proves the inclusion $\mathcal{W}_q \subseteq \mathcal{W}_p$, and the converse inclusion can be shown analogously. As X is path-connected, there are $p = p_1, p_2, \ldots, p_k = q \in X$ with $p_1 \models^* p_2, p_2 \models^* p_3, \ldots, p_{k-1} \models^* p_k$. We show by induction that $C \in \mathcal{W}_{p_j}$ for all $j = 1, 2, \ldots, k$. If $j = 1$, then $C \in C_{p1}$ by $p_1 = p$. Now let $1 \leq j < k$ and assume $C \in \mathcal{W}_{p_j}$. By $p_j \models^* p_{j+1}$ there is a set $Y \subseteq X$ such that $\{p_j\} \cup Y$ and $\{\neg p_{j+1}\} \cup Y$ are each consistent but $\{p_j, p_{j+1}\} \cup Y$ is inconsistent. It follows that each of $\{p_j, p_{j+1}\} \cup Y$ and $\{\neg p_j, \neg p_{j+1}\} \cup Y$ is consistent (using L3 in conjunction with L1, L2). So we may define a profile $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ by

$$A_i := \begin{cases} A_{\{p_j, p_{j+1}\} \cup Y} & \text{if } i \in C \\ A_{\{\neg p_j, \neg p_{j+1}\} \cup Y} & \text{if } i \in N \backslash C. \end{cases}$$

Since $Y \subseteq A_i$ for all i, $Y \subseteq F\{A_1, \ldots, A_n\}$ by claim 1. Since $\{i : p_j \in A_i\} = C \in \mathcal{W}_{p_j}$, we have $p_j \in F(A_1, \ldots, A_n)$. So $\{p_j\} \cup Y \subseteq F(A_1, \ldots, A_n)$. Hence, since $\{p_j, \neg p_{j+1}\} \cup Y$ is inconsistent, $\neg p_{j+1} \notin F(A_1, \ldots, A_n)$, whence $p_{j+1} \in F(A_1, \ldots, A_n)$. So, as $\{i : p_{j+1} \in A_i\} = C$, we have $C \in \mathcal{W}_{p_{j+1}}$, as desired.

As $\mathcal{W}_p$ is the same set for each contingent $p \in X$, let $\mathcal{W}$ be this set. To complete the proof of the claim, it is sufficient to show that, for every $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ and every $p \in X$, $p \in F(A_1, \ldots, A_n)$ if and only if $\{i : p \in A_i\} \in \mathcal{W}$. If p is contingent this holds by definition of $\mathcal{W}$; if p is a tautology it holds because $p \in F(A_1, \ldots, A_n)$ (by collective rationality), $\{i : p \in A_i\} = N$ (by universal domain) and $N \in \mathcal{W}$ (by claim 1); analogously, if p is a contradiction it holds because $p \notin F(A_1, \ldots, A_n)$, $\{i : p \in A_i\} = \emptyset$ and $\emptyset \notin \mathcal{W}$.

Claim 3. (1) $N \in \mathcal{W}$; (2) for every coalition $C \subseteq N$, $C \in \mathcal{W}$ if and only if $N \backslash C \notin \mathcal{W}$; (3) for every coalitions C, $C^* \subseteq N$, if $C \in \mathcal{W}$ and $C \subseteq C^*$ then $C^* \in \mathcal{W}$.

Part (1) follows from claim 1. Regarding parts (2) and (3), note that, for any $C \subseteq N$, there exists a $p \in X$ and an $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ with $\{i : p \in A_i\} = C$; this holds because X contains a contingent proposition p. Part (2) holds because, for any $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$, each of the sets $A_1, \ldots, A_n$, $F(A_1, \ldots, A_n)$ contains exactly one member of any pair $p, \neg p \in X$, by universal domain and collective rationality. Part (3) follows from a repeated application of monotonicity and universal domain.

Claim 4. There exists an inconsistent set $Y \subseteq X$ with pairwise disjoint subsets Z_1, Z_2, Z_3 such that $(Y \backslash Z_j) \cup Z_j^\neg$ is consistent for any $j \in \{1, 2, 3\}$. Here, $Z^\neg := \{\neg p : p \in Z\}$ for any $Z \subseteq X$.

By assumption, there exists a contingent $p \in X$; also $\neg p$ is then contingent. So, by path-connectedness, there exist $p = p_1, p_2, \ldots, p_k = \neg p \in X$ and $Y_1^*, Y_2^*, \ldots, Y_{k-1}^* \subseteq X$ such that

(3) for each $t \in \{1, \ldots, k-1\}$, $\{p_t, \neg p_{t+1}\} \cup Y_t^*$ is inconsistent; and

(4) for each $t \in \{1, \ldots, k-1\}$, $\{p_t\} \cup Y_t^*$ and $\{\neg p_{t+1}\} \cup Y_t^*$ are consistent.

From (3) and (4) it follows (using L3 in conjunction with L1, L2) that

(5) for each $t \in \{1, \ldots, k-1\}$, $\{p_t, p_{t+1}\} \cup Y_t^*$ and $\{\neg p_t, \neg p_{t+1}\} \cup Y_t^*$ are consistent.

We first show that there exists a $t \in \{1, \ldots, k-1\}$ such that $\{p_t, \neg p_{t+1}\}$ is consistent. Assume for contradiction that each of $\{p_1, \neg p_2\}, \ldots, \{p_{k-1}, \neg p_k\}$ is inconsistent. So (using L2) each of $\{p_1, \neg p_2\}, \{p_1, p_2, \neg p_3\}, \ldots, \{p_1, \ldots, p_{k-1}, \neg p_k\}$ is inconsistent. As $\{p_1\} = \{p\}$ is consistent, either $\{p_1, p_2\}$ or $\{p_1, \neg p_2\}$ is consistent (by L2 and L3); hence, as $\{p_1, \neg p_2\}$ is inconsistent, $\{p_1, p_2\}$ is consistent. So either $\{p_1, p_2, p_3\}$ or $\{p_1, p_2, \neg p_3\}$ is consistent (again by L2 and L3); hence, as $\{p_1, p_2, \neg p_3\}$ is inconsistent, $\{p_1, p_2, p_3\}$ is consistent. Continuing this argument, it follows after $k-1$ steps that $\{p_1, \ldots, p_k\}$ is consistent. Hence $\{p_1, p_k\}$ is consistent (by L2), i.e., $\{p, \neg p\}$ is consistent, a contradiction (by L1).

We have shown that there is a $t \in \{1, \ldots, k-1\}$ such that $\{p_t, \neg p_{t+1}\}$ is consistent, whence $Y_t^* \neq \emptyset$ by (3). Define $Y := \{p_t, \neg p_{t+1}\} \cup Y_t^*$, $Z_1 := \{p_t\}$, and $Z_2 := \{\neg p_{t+1}\}$. Since $\{p_t, \neg p_{t+1}\}$ is consistent, $\{p_t, \neg p_{t+1}\} \cup B$ is consistent for some set B that contains q or $\neg q$ (but not both) for each $q \in Y_t^*$ (by L3 together with L1, L2). Note that there exists a $Z_3 \subseteq Y_t^*$ with $B = (Y_t^* \backslash Z_3) \cup Z_3^{\neg}$. This proves the claim, since:

- $Y = \{p_t, \neg p_{t+1}\} \cup Y_t^*$ is inconsistent by (3),
- Z_1, Z_2, Z_3 are pairwise disjoint subsets of Y,
- $(Y \backslash Z_1) \cup Z_1^{\neg} = (Y \backslash \{p_t\}) \cup \{\neg p_t\} = \{\neg p_t, \neg p_{t+1}\} \cup Y_t^*$ is consistent by (4),
- $(Y \backslash Z_2) \cup Z_2^{\neg} = (Y \backslash \{\neg p_{t+1}\}) \cup \{p_{t+1}\} = \{p_t, p_{t+1}\} \cup Y_t^*$ is consistent by (4),
- $(Y \backslash Z_3) \cup Z_3^{\neg} = \{p_t, \neg p_{t+1}\} \cup (Y_t^* \backslash Z_3) \cup Z_3^{\neg} = \{p_t, \neg p_{t+1}\} \cup B$ is consistent.

Claim 5. For any coalitions $C, C^* \subseteq N$, if $C, C^* \in \mathcal{W}$ then $C \cap C^* \in \mathcal{W}$.

Consider any $C, C^* \in \mathcal{W}$, and assume for contradiction that $C_1 := C \cap C^* \notin \mathcal{W}$. Put $C_2 := C^* \backslash C$ and $C_3 := N \backslash C^*$. Let Y, Z_1, Z_2, Z_3 be as in claim 4. Noting that C_1, C_2, C_3 form a partition of N, we define the profile $(A_1, \ldots, A_n)$ by:

$$A_i := \begin{cases} A_{(Y \backslash Z_1) \cup Z_1^{\neg}} & \text{if } i \in C_1 \\ A_{(Y \backslash Z_2) \cup Z_2^{\neg}} & \text{if } i \in C_2 \\ A_{(Y \backslash Z_3) \cup Z_3^{\neg}} & \text{if } i \in C_3. \end{cases}$$

By $C_1 \notin \mathcal{W}$ and $N \backslash C_1 = C_2 \cup C_3$ we have $C_2 \cup C_3 \in \mathcal{W}$ by claim 3, and so $Z_1 \subseteq F(A_1, \ldots, A_n)$. By $C \in \mathcal{W}$ and $C \subseteq C_1 \cup C_3$ we have $C_1 \cup C_3 \in \mathcal{W}$ by claim 3, and so $Z_2 \subseteq F(A_1, \ldots, A_n)$. Further, $Z_3 \subseteq F(A_1, \ldots, A_3)$ as $C_1 \cup C_2 = C^* \in \mathcal{W}$. Finally, $Y \backslash (Z_1 \cup Z_2 \cup Z_3) \subseteq F(A_1, \ldots, A_n)$ as $N \in \mathcal{W}$ by claim 3. In summary, we have $Y \subseteq F(A_1, \ldots, A_n)$, violating consistency.

Claim 6. There is a dictator.

Consider the intersection of all winning coalitions, $\tilde{C} := \cap_{C \in \mathcal{W}} C$. By claim 5, $\tilde{C} \in \mathcal{W}$. So $\tilde{C} \neq \emptyset$, as by claim 3, $\emptyset \notin \mathcal{W}$. Hence there is a $j \in \tilde{C}$. To show that j is a dictator, consider any $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$ and $p \in X$, and let us prove that $p \in F(A_1, \ldots, A_n)$ if and only if $p \in A_j$. If $p \in F(A_1, \ldots, A_n)$ then $C := \{i : p \in A_i\} \in \mathcal{W}$, whence $j \in C$ (as j belongs to every winning coalition), i.e., $p \in A_j$. Conversely, if $p \notin F(A_1, \ldots, A_n)$, then $\neg p \in F(A_1, \ldots, A_n)$; so by an argument analogous to the previous one, $\neg p \in A_j$, whence $p \notin A_j$.

Proof of Theorem 4. Let $Y \subseteq X$.

(i) First, assume F is strategy-proof for C_Y. To show non-manipulability on Y, consider any proposition $p \in Y$, individual i, and profile $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$, such that $F(A_1, \ldots, A_n)$ disagrees with A_i

on p. Let $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$ be any i-variant. We have to show that $F(A_1, \ldots, A_i^*, \ldots, A_n)$ still disagrees with A_i on p. Define a preference relation $\succsim_i$ over judgment sets by [$B \succsim_i B^*$ if and only if A_i agrees on p with B but not with B^*, or with both B and B^*, or with neither B nor B^*]. ($\succsim_i$ is interpreted as individual i's preference relation in case i cares only about p.) It follows immediately that $\succsim_i$ is reflexive and transitive and respects closeness to A_i on Y, i.e., is a member of $C_Y(A_i)$. So, by strategy-proofness for C_Y, $F(A_1, \ldots, A_n) \succsim_i F(A_1, \ldots, A_i^*, \ldots, A_n)$. Since A_i disagrees with $F(A_1, \ldots, A_n)$ on p, the definition of $\succsim_i$ implies that A_i still disagrees with $F(A_1, \ldots, A_i^*, \ldots, A_n)$ on p.

(ii) Now assume that F is non-manipulable on Y. To show strategy-proofness for C_Y, consider any individual i, profile $(A_1, \ldots, A_n) \in Domain(F)$, and preference relation $\succsim_i \in C_Y(A_i)$, and let $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$ be any i-variant. We have to prove that $F(A_1, \ldots, A_n) \succsim_i F(A_1, \ldots, A_i^*, \ldots, A_n)$. By non-manipulability on Y, for every proposition $p \in Y$, if A_i disagrees with $F(A_1, \ldots, A_n)$ on p, then also with $F(A_1, \ldots, A_i^*, \ldots, A_n)$; in other words, if A_i, agrees with $F(A_1, \ldots, A_i^*, \ldots, A_n)$ on p, then also with $F(A_1, \ldots, A_n)$. So $F(A_1, \ldots, A_n)$ is at least as close to A_i on Y as $F(A_1, \ldots, A_i^*, \ldots, A_n)$. Hence $F(A_1, \ldots, A_n) \succsim_i F(A_1, \ldots, A_i^*, \ldots, A_n)$, as $\succsim_i \in C_Y(A_i)$.

Proof of Proposition 1. We prove this result directly, although it can also be derived from Corollary 1. Let F be premise-based voting. To show that F is strategy-proof for $C_{Y_{reason}}$, consider any individual i, profile $(A_1, \ldots, A_n) \in Domain(F)$, i-variant $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$, and preference relation $\succsim_i \in C_{Y_{reason}}(A_i)$. The definition of premise-based voting implies that $F(A_1, \ldots, A_n)$ is at least as close to A_i as $F(A_1, \ldots, A_i^*, \ldots, A_n)$ on Y_{reason}. So, by $\succsim_i \in C_{Y_{reason}}(A_i)$, we have $F(A_1, \ldots, A_n) \succsim_i F(A_1, \ldots, A_i^*, \ldots, A_n)$.

Proof of Theorem 5. Consider the conjunctive agenda (the proof is analogous for disjunctive agendas). Let F and G be premise- and conclusion-based voting, respectively. Take any profile $(A_1, \ldots, A_n) \in Domain(F) = Domain(G)$ and any preference relations $\succsim_i \in C_{Y_{outcome}}(A_1), \ldots, \succsim_n \in C_{Y_{outcome}}(A_n)$. Define $(B_1, \ldots, B_n)$ by

$$B_i = \begin{cases} \{\neg a_1, \ldots, \neg a_k, c \leftrightarrow (a_1 \wedge \cdots \wedge a_k), \neg c\} & \text{if } \neg c \in A_i, \\ \{a_1, \ldots, a_k, c \leftrightarrow (a_1 \wedge \cdots \wedge a_k), c\} & \text{if } c \in A_i. \end{cases}$$

It can easily be seen that, for each i and any pair of i-variants $(D_i, \ldots, B_i, \ldots, D_n)$, $(D_1, \ldots, B_1^*, \ldots, D_n) \in Domain(F)$, $F(D_1, \ldots, B_i, \ldots, D_n)$ is at least as close to A_i on $Y_{outome}(= \{c, \neg c\})$ as $F(D_1, \ldots, B_i^*, \ldots, D_n)$; so $(D_1, \ldots, B_i, \ldots, D_n) \succsim_i (D_1, \ldots, B_i^*, \ldots, D_n)$ as $\succsim_i \in C_{Y_{outcome}}(A_i)$. Hence, submitting B_i is a weakly dominant strategy for each i under F. Second,

let $(C_1, \ldots, C_n)$ be $(A_1, \ldots, A_n)$ (the truthful profile). Then, for each i, submitting C_i is a weakly dominant strategy under G, as G is strategy-proof. Finally, it can easily be seen that $F(B_1, \ldots, B_n)$ and $G(C_1, \ldots, C_n) = G(A_1, \ldots, A_n)$ agree on each proposition in $Y_{outcome} = \{c, \neg c\}$.

REFERENCES

Barberà, S., F. Gul and E. Stacchetti. 1993. Generalized Median Voter Schemes and Committees. *Journal of Economic Theory* 61: 262–89.

Barberà, S., J. Massóa and A. Nemeb. 1997. Voting under constraints. *Journal of Economic Theory* 76(2): 298–321.

Baigent, N. 1987. Preference proximity and anonymous social choice. *Quarterly Journal of Economics* 102(1): 161–70.

Bossert, W., and T. Storcken. 1992. Strategy-proofness of social welfare functions: the use of the Kemeny distance between preference orderings. *Social Choice and Welfare* 9: 345–60.

Bovens, L., and W. Rabinowicz. 2006. Democratic answers to complex questions–an epistemic perspective. *Synthese* 150: 131–53.

Brams, S. J., D. M. Kilgour and W. S. Zwicker. 1997. Voting on referenda: the separability problem and possible solutions. *Electoral Studies* 16(3): 359–77

Brams, S.J., D. M. Kilgour and W. S. Zwicker. 1998. The paradox of multiple elections. *Social Choice and Welfare* 15: 211–36.

Brennan, G. 2001. Collective Coherence? *International Review of Law and Economics* 21: 197–211.

Chapman, B. 1998. More easily done than said: Rules, reason and rational social choice. *Oxford Journal of Legal Studies* 18: 293–330.

Chapman, B. 2002. Rational Aggregation. *Politics, Philosophy and Economics* 1: 337–54.

Dietrich, F. 2006. Judgment Aggregation: (Im)Possibility Theorems. *Journal of Economic Theory* 126: 286–98.

Dietrich, F. 2007. A generalised model of judgment aggregation. *Social Choice and Welfare* 28(4): 529–65.

Dietrich F. Forthcoming. The possibility of judgment aggregation on agendas with subjunctive implications. *Journal of Economic Theory*.

Dietrich, F., and C. List. 2007a. Arrow's theorem in judgment aggregation. *Social Choice and Welfare* 29(1): 19–33.

Dietrich, F., and C. List. 2007b. Judgment aggregation by quota rules. *Journal of Theoretical Politics* 19(4, in press).

Dokow, E., and R. Holzman. 2005. *Aggregation of binary evaluations*. Working paper, Technion Israel Institute of Technology.

Dryzek, J., and C. List. 2003. Social choice theory and deliberative democracy: A reconciliation. *British Journal of Political Science* 33: 1–28.

Elster, J. 1986. The Market and the forum. In *Foundations of Social Choice Theory*, ed. J. Elster and A. Hylland. Cambridge, Cambridge University Press, 103–32.

Ferejohn, J. 2003. *Conversability and collective intention*. Paper presented at the Common Minds Conference, Australian National University, 24–25 July 2003.

Gärdenfors, P. 2006. An Arrow-like theorem for voting with logical consequences. *Economics and Philosophy* 22(2): 181–90.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41(July): 587–601.

Goodin, R. E. 1986. Laundering preferences. In *Foundations of Social Choice Theory*, ed. J. Elster and A. Hylland. Cambridge, Cambridge University Press, 75–101.

Grofman, B. 1985. Research note: The accuracy of group majorities for disjunctive and conjunctive decision tasks. *Organizational Behavior and Human Decision Processes* 35: 119–23.

van Hees, M. 2007. The limits of epistemic democracy. *Social Choice and Welfare* 28(4): 649–66.
Kelly, J. S. 1989. The Ostrogorski Paradox. *Social Choice and Welfare* 6: 71–6.
Konieczny, S. and R. Pino-Perez. 2002. Merging information under constraints: a logical framework. *Journal of Logic and Computation* 12: 773–808.
Kornhauser, L. A. and L. G. Sager. 1986. Unpacking the Court. *Yale Law Journal* 96: 82–117.
List, C. 2002a. Two concepts of agreement. *The Good Society* 11(1): 72–9.
List, C. 2002b. Discursive path-dependencies. Nuffield College Working Paper in Politics 2002-W15 (9 May 2002).
List, C. 2003. A Possibility Theorem on Aggregation over Multiple Interconnected Propositions. *Mathematical Social Sciences* 45: 1–13 (with Corrigendum in *Mathematical Social Sciences* 52: 109–10).
List, C. 2004. A model of path dependence in decisions over multiple propositions. *American Political Science Review* 98: 495–513.
List, C. 2005. The probability of inconsistencies in complex collective decisions. *Social Choice and Welfare* 24: 3–32.
List, C. 2006. The discursive dilemma and public reason. *Ethics* 116: 362–402.
List, C. and P. Pettit. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy* 18: 89–110.
List, C. and P. Pettit. 2004. Aggregating sets of judgments: Two impossibility results compared. *Synthese* 140(1–2): 207–35.
Miller, D. 1992. Deliberative democracy and social choice. *Political Studies* 40: 54–67.
Nehring, K. 2003. Arrow's theorem as a corollary. *Economics Letters* 80: 379–82.
Nehring, K. and C. Puppe. 2002. *Strategyproof social choice on single-peaked domains: Possibility, impossibility and the space between*. Working paper, University of California at Davis.
Nehring, K. and C. Puppe. 2005. Consistent judgement aggregation: A characterization. Working paper, University of Karlsruhe.
Osherson, D. and M. Vardi. Forthcoming. Aggregating disparate estimates of chance. *Games and Economic Behavior*.
Pauly, M. and M. van Hees. 2006. Logical constraints on judgment aggregation. *Journal of Philosophical Logic* 35: 569–85.
Pettit, P. 2001. Deliberative democracy and the discursive dilemma. *Philosophical Issues* 11: 268–99.
Pigozzi, G. 2006. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese* 152(2): 285–98.
Saporiti, A. and F. Thomé. 2005. *Strategy-proofness and single-crossing*. Working paper, Queen Mary, University of London.
Satterthwaite, M. 1975. Strategyproofness and Arrow's conditions: existence and correspondences for voting procedures and social welfare functions. *Journal of Economic Theory* 10: 187–217.
Schulte, O. 2005. Minimal belief change, Pareto-optimality and logical consequence. *Economic Theory* 19(1): 105–44.
Sunstein, C. 1994. *Political Conflict and Legal Agreement*. Tanner Lectures on Human Values, Harvard.
Taylor, A. D. 2002. The Manipulability of Voting Systems. American Mathematical Monthy. 109: 321–37.
Taylor, A. D. 2005. *Social Choice and the Mathematics of Manipulation*. Cambridge, Cambridge University Press.
Wilson, R. 1975. On the theory of aggregation. *Journal of Economic Theory* 10: 89–99.