Strategy-proof judgment aggregation

Franz Dietrich
(University of Konstanz)

and

Christian List
(LSE)

Political Economy and Public Policy Series
The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

PEPP/9
July 2005

Tel: (020) 7955 6674

Strategy-proof judgment aggregation

Franz Dietrich and Christian List[1]

In the theory of judgment aggregation on logically connected propositions, an important question remains open: Which aggregation rules are manipulable and which are strategy-proof? We define manipulability and strategy-proofness in judgment aggregation, characterize all strategy-proof aggregation rules, and prove an impossibility theorem similar to the Gibbard-Satterthwaite theorem. Among other escape-routes from the impossibility, we discuss weakening strategy-proofness itself. Comparing two prominent aggregation rules, we show that conclusion-based voting is strategy-proof, but generates incomplete judgments, while premise-based voting is only strategy-proof 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.

Keywords: Judgment aggregation, strategy-proofness, logic, Gibbard-Satterthwaite theorem

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? This problem – *judgment aggregation* – is discussed in a growing literature and generalizes earlier problems of social choice, notably preference aggregation in the Condorcet-Arrow tradition.[2] Judgment aggregation is often illustrated by a paradox: the *discursive* (or *doctrinal*) *paradox* (Kornhauser and Sager [17]; Brennan [5]; Pettit [27]; Bovens and Rabinowicz [3]). Suppose a university committee responsible for a tenure decision has to make collective judgments on three propositions:

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.

	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

Table 1: The discursive paradox

[1]F. Dietrich, ZWN, University of Konstanz, 78457 Konstanz, Germany, franz.dietrich@uni-konstanz.de; C. List, Dept. of Govt., LSE, London WC2A 2AE, U.K., c.list@lse.ac.uk. This paper was presented at Konstanz (6/2004), the Osaka SCW Conference (7/2004), LSE (10/2004), Université de Caen (11/2004), UEA (1/2005), Northwestern University (5/2005), the Vigo SAET 2005 Conference (6/2005). We thank the participants at these occasions for comments. Revised 7/2005.

[2]Preference aggregation becomes a case of judgment aggregation by expressing preference relations as sets of binary ranking propositions in predicate logic (List and Pettit [22], Dietrich and List [11]).

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 (Pettit [27]; Chapman [6]; Bovens and Rabinowicz [3]; List [19]). 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. 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.

Abstracting from the discursive dilemma, List and Pettit ([21], [22]) 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. Stronger impossibility results have been proved by Pauly and van Hees [26], Dietrich [7], Gärdenfors [14], Nehring and Puppe [23] and Dokow and Holzman [12], and possibility results by List ([18], [20]), Dietrich [7] and Pigozzi [28]. Judgment aggregation in multi-valued logics and general logics, respectively, has been analysed by Pauly and van Hees ([26]; also van Hees [16]) and Dietrich [8]. Probabilistic judgment aggregation has been analysed by Osherson and Vardi [25].

But one important question has not been investigated yet: Which judgment aggregation rules are manipulable by strategic voting and which are strategy-proof? The answer to this question is not obvious, as strategy-proofness is a preference-theoretic concept and preferences are not primitives of judgment aggregation models. Yet the question matters for the design and implementation of aggregation rules. Ideally, we would like to find rules that lead individuals to reveal their judgments truthfully.

Here we aim to fill this gap in the literature. We first introduce a simple condition of non-manipulability and characterize the class of non-manipulable judgment aggregation rules. We then show that our condition is equivalent to a game-theoretically motivated strategy-proofness condition similar to the one introduced by Gibbard [15] and Satterthwaite [30] for preference aggregation (for recent work in different informational frameworks from ours, see Barberà et al. [1], Nehring and Puppe [24], Saporiti and Thomé [29]). Our characterization of non-manipulable aggregation rules yields a characterization of strategy-proof aggregation rules.[3]

We prove that, for a general class of aggregation problems including the tenure example above, there exists no strategy-proof judgment aggregation rule satisfying universal domain and some other mild conditions, an impossibility result similar to the Gibbard-Satterthwaite theorem on preference aggregation.

In addition to identifying escape-routes from the impossibility, we show that our default condition of strategy-proofness falls into a general family of conditions and discuss weaker conditions in this family. In the tenure example, conclusion-based voting is strategy-proof, but produces no collective judgments on the premises. Premise-based voting satisfies only the weaker condition of strategy-proofness for

[3] List [20] has stated sufficient but not necessary conditions for strategy-proofness in *sequential* judgment aggregation.

"reason-oriented" individuals, as defined below. 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 of broader interest as premise-based voting has been advocated, and conclusion-based voting rejected, in debates on "deliberative democracy" (Pettit [27]).

2 The basic model

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

2.1 Formal logic

Propositions can be represented in any logic satisfying some minimal conditions (Dietrich [8], Dietrich and List [10]). We use standard propositional logic in our examples, but our results do not require this restriction.

A *logic (with negation symbol $\neg$)* is defined by a non-empty set $\mathbf{L}$ of *propositions* (where $p \in \mathbf{L}$ implies $\neg p \in \mathbf{L}$) and an *entailment relation* $\vDash$ (where, for each $A \subseteq \mathbf{L}$ and each $p \in \mathbf{L}$, $A \vDash p$ is read as "A entails p").[4]

A set $A \subseteq \mathbf{L}$ is *inconsistent* if $A \vDash p$ and $A \vDash \neg p$ for some $p \in \mathbf{L}$, and *consistent* otherwise. A proposition $p \in \mathbf{L}$ is *contingent* if both $\{p\}$ and $\{\neg p\}$ are consistent. We require our logic to satisfy the following ($p \vDash q$ stands for $\{p\} \vDash q$):

- (L1) For all $p \in \mathbf{L}$, $p \vDash p$ (self-entailment).
- (L2) For all $p \in \mathbf{L}$ and $A \subseteq B \subseteq \mathbf{L}$, if $A \vDash p$ then $B \vDash p$ (monotonicity).
- (L3) $\emptyset$ is consistent, and each consistent set $A \subseteq \mathbf{L}$ has a consistent superset $B \subseteq \mathbf{L}$ containing a member of each pair $p, \neg p \in \mathbf{L}$ (completability).

Apart from standard propositional logic, many different logics satisfy these conditions, including predicate, modal, conditional and deontic logics; non-compact or paraconsistent logics are also admissible.[5]

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 $\{p, \neg p\}$

[4] $\vDash$ 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.

[5] The aggregation problem introduced by Wilson [32] and revisited by Dokow and Holzman [12], where a group has to determine its yes/no views on several issues based on the group members' views on these issues (subject to feasibility constraints), can also be embedded into our model (under this embedding, Dokow and Holzman's results apply to a logic satisfying L1 to L3 and compactness or a finite agenda). Hence our analysis of manipulation and strategy-proofness applies to Wilson's model too. Unlike our model, Wilson's does not permit a general representation of an entailment relation (and an analysis of deductive closure), as its primitive is a notion of consistency (feasibility), from which an entailment relation can be retrieved only for certain logics (Dietrich [8]). A similar remark applies to the treatment of feasibility in Barberà et al. [1] and Nehring and Puppe [24].

(with p a non-negated proposition). For simplicity, we assume that double negations cancel each other out, i.e. $\neg\neg p$ stands for p.[6]

Two important examples are *conjunctive* and *disjunctive* agendas in standard propositional logic.[7] A conjunctive agenda is $X := \{a_1, \neg a_1, \ldots, a_k, \neg a_k, c, \neg c, c \leftrightarrow (a_1 \wedge \cdots \wedge a_k), \neg(c \leftrightarrow (a_1 \wedge \cdots \wedge a_k))\}$, 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. In the tenure example above, we have a conjunctive agenda with $k = 2$. To define a disjunctive agenda, we replace $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$ with $c \leftrightarrow (a_1 \vee \cdots \vee a_k)$.

Other examples of agendas are the *simple network agenda* (in standard propositional logic or a suitable conditional logic) $X := \{a, \neg a, b, \neg b, a \rightarrow b, \neg(a \rightarrow b)\}$ and the *Arrow agenda* $X := \{xRy, \neg xRy : x, y \in K\}$, where the underlying logic is a simple predicate logic with a set of constants K representing options ($|K| \geq 3$) and a two-place predicate R representing preferences, as defined in Dietrich and List [10].[8]

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. Our main impossibility theorem holds for a large class of agendas, defined below. We also discuss some applications to special 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. A judgment set A_i is *consistent* if it is a consistent set of propositions as defined above; A_i is *complete* if, for every proposition $p \in X$, $p \in A_i$ or $\neg p \in A_i$. 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.

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

2.4 Examples of aggregation rules

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

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

[6] Hereafter we use $\neg$ to represent a modified negation symbol $\sim$, where $\sim p := \neg p$ if p is unnegated and $\sim p := q$ if $p = \neg q$ for some q.

[7] Here $\mathbf{L}$ is the smallest set such that (i) $\mathbf{L}$ contains the given *atomic* propositions a, b, c, ..., and (ii) if $\mathbf{L}$ contains two propositions p and q, then $\mathbf{L}$ also contains $\neg p$, $(p \wedge q)$, $(p \vee q)$, $(p \rightarrow q)$, $(p \leftrightarrow q)$, with the *connectives* $\neg$ (not), $\wedge$ (and), $\vee$ (or), $\rightarrow$ (if-then), $\leftrightarrow$ (if and only if). We drop brackets when there is no ambiguity; for example, we write $a \rightarrow (b \wedge c \wedge d)$ instead of $(a \rightarrow (b \wedge (c \wedge d)))$. The entailment relation $\models$ is defined in the standard way.

[8] The entailment relation $\models$ in this logic 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 $\{(\forall v)vRv$, $(\forall v_1)(\forall v_2)(\forall v_3)((v_1Rv_2 \wedge v_2Rv_3) \rightarrow v_1Rv_3)$, $(\forall v_1)(\forall v_2)(\neg v_1 = v_2 \rightarrow (v_1Rv_2 \vee v_2Rv_1))\}$.

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

Premise-based voting. For each $(A_1, \ldots, A_n)$, $F(A_1, \ldots, A_n)$ 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 $p \notin F(A_1, \ldots, A_n)$.

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

Conclusion-based voting. For each $(A_1, \ldots, A_n)$, $F(A_1, \ldots, A_n)$ 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 deliberative democracy and the discursive paradox, several arguments have been offered for the superiority of premise-based voting over conclusion-based voting.[9] 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 simple condition of non-manipulability, not yet explicitly game-theoretic. Below we prove its equivalence to a game-theoretically motivated strategy-proofness condition.

3.1 An example

Consider the profile in Table 1 again. 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

[9] One such argument draws on a "deliberative" conception of democracy, which emphasizes that collective decisions on conclusions should follow from collectively decided premises (Pettit [27]). 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 (Bovens and Rabinowicz [3], List [19]).

is the final tenure decision, not the underlying reasons; they are "outcome-oriented", as defined precisely later.

Suppose first that 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 that 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 [13] 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 $(A_1, \ldots, A_i^*, \ldots, A_n) \in Domain(F)$.

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 discussed later when we introduce preferences over judgment sets.[10]

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

[10] Individuals may or may not act on *opportunities* for manipulation. Whether they have *incentives* to act on such opportunities depends on how much they care about the propositions involved in a possible act of manipulation.

Non-manipulability on Y. F is not manipulable at any profile by any individual on any proposition in Y. Equivalently, for every 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 $Y \subseteq X$; 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, using an independence condition and a monotonicity condition. Let $Y \subseteq X$.

Independence on Y. For every 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 every 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 every 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 *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, (ii) and (iii) are equivalent, and each implies (i).

Theorem 1 holds for any agenda. Also, no assumption on the consistency or completeness of collective judgments is needed. 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 premise-based voting is independent and monotonic on the set of premises $Y = \{a_1, \neg a_1, \ldots, a_k, \neg a_k\}$, hence non-manipulable on Y.

Proof. 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 $(A_1, ..., A_n) \in Domain(F)$, such that $F(A_1, ..., A_n)$ disagrees with A_i on p. Take any i-variant $(A_1, ..., A_i^*, ..., A_n) \in Domain(F)$. We have to show that $F(A_1, ..., A_i^*, ..., A_n)$ still disagrees with A_i on p. Assume first that A_i and A_i^* agree on p. Then in both profiles $(A_1, ..., A_n)$ and $(A_1, ..., A_i^*, ..., A_n)$ exactly the same individuals accept p. Hence, by independence on Y, $F(A_1, ..., A_i^*, ..., A_n)$ agrees with F $(A_1, ..., A_n)$ on p, hence disagrees with A_i on p. Now assume A_i^* disagrees with A_i on p, i.e. agrees with $F(A_1, ..., A_n)$ on p. Then, by monotonicity on Y, $F(A_1, ..., A_i^*, ..., A_n)$ agrees with $F(A_1, ..., A_n)$ 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 $(A_1, ..., A_n), (A_1, ..., A_i^*, ..., A_n) \in Domain(F)$ with $p \notin A_i$ and $p \in A_i^*$. If $p \in F(A_1, ..., A_n)$, then A_i disagrees on p with $F(A_1, ..., A_n)$, hence also with $F(A_1, ..., A_i^*, ..., A_n)$ by non-manipulability on Y. So $p \in F(A_1, ..., A_i^*, ..., A_n)$. To show independence on Y, consider any proposition $p \in Y$ and profiles $(A_1, ..., A_n)$, $(A_1^*, ..., A_n^*) \in Domain(F)$ such that [for all individuals i, $p \in A_i$ if and only if $p \in A_i^*$]. In other words, for all individuals i, A_i and A_i^* agree on p. We have to show that $F(A_1, ..., A_n)$ and $F(A_1^*, ..., A_n^*)$ agree on p. Starting with the profile $(A_1, ..., A_n)$, we replace first A_1 by A_1^*, then A_2 by A_2^*, ..., 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^* disagrees with the respective outcome (but not both). 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 $F(A_1, ..., A_n)$ and $F(A_1^*, ..., A_n^*)$ agree on p, which proves independence on Y. ■

3.4 An impossibility result

Ideally, we want to achieve non-manipulability *simpliciter* and not just on some subset $Y \subsetneq X$. Conclusion-based voting is non-manipulable in this strong sense, but generates incomplete collective judgments. Are there any non-manipulable aggregation rules that generate complete and consistent 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.

We write $p \vDash^* q$ when $Y \cup p \vDash q$ for some set $Y \subseteq X$ consistent with p and with $\neg q$. An agenda X is *path-connected* if, for any contingent propositions $p, q \in X$, there is a sequence $p_1, p_2, ..., p_k \in X$ $(k \geq 1)$ with $p = p_1$ and $q = p_k$ such that $p_1 \vDash^* p_2, p_2 \vDash^* p_3, ..., p_{k-1} \vDash^* p_k$ (Dietrich and List [10]; for the related notion of *total-blockedness*, see Nehring and Puppe [23]).[11] We show in the Appendix that conjunctive and disjunctive agendas fall into the class of path-connected agendas. The Arrow agenda, as defined above, is also path-connected (Dietrich and List [10]). Consider the following conditions in addition to universal domain.

Collective Rationality. For every profile $(A_1, \ldots, A_n) \in \mathit{Domain}(F)$, $F(A_1, \ldots, A_n)$ is complete and consistent.

Responsiveness. For every contingent proposition $p \in X$, there exist two profiles $(A_1, \ldots, A_n)$, $(A_1^*, \ldots, A_n^*) \in \mathit{Domain}(F)$ 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 (e.g. a conjunctive, disjunctive or Arrow 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.*

In the case of a compact logic, the result could also be derived from Theorem 1 and Nehring and Puppe's [23] impossibility theorem on monotonic and independent aggregation rules for totally blocked agendas. Below we restate the impossibility result of Theorem 2 using a game-theoretically motivated strategy-proofness condition. Our result 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 [15], Satterthwaite [30]). In the special case of the Arrow agenda, however, there is an interesting disanalogy between Theorem 2 and the Gibbard-Satterthwaite theorem. As a collectively rational judgment aggregation rule for the Arrow agenda represents an Arrowian social welfare function, Theorem 2 establishes an impossibility result on the non-manipulability of social welfare functions (generating orderings) as opposed to social choice functions (generating winning options); for a related result, see Bossert and Storcken [4].

[11] Our results and proofs also hold for a weaker (but less easily readable) definition of path-connectedness, obtained by defining $p \vDash^* q$ if and only if $Y \cup \{p, \neg q\}$ is inconsistent for some $Y \subseteq X$ consistent with p and with $\neg q$. The two definitions are equivalent for non-paraconsistent logics. The weaker definition is also weaker than Nehring and Puppe's *total-blockedness*, which is logically unrelated to the stronger definition.

Proof. 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 $(A_1, ..., A_n) \in Domain(F)$, if $p \in A_i$ for each i then $p \in F(A_1, ..., A_n)$.

Consider any $p \in X$ and $(A_1, ..., A_n) \in Domain(F)$ 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), $p \in F(A_1, ..., A_n)$ 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 $(B_1, ..., B_n) \in Domain(F)$ such that $p \in F(B_1, ..., B_n)$. In $(B_1, ..., B_n)$ we now replace one by one each judgment set B_i by A_i, until we obtain the profile $(A_1, ..., A_n)$. 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 $p \in F(A_1, ..., A_n)$, as desired.

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

For each $p \in X$, let $\mathcal{W}_p$ be the set all subsets $C \subseteq N$ such that $p \in F(A_1, ..., A_n)$ for some (hence by independence any) $(A_1, ..., A_n) \in Domain(F)$ with $\{i : p \in A_i\} = C$. Note that $F(A_1, ..., A_n) = \{p \in X : \{i : p \in A_i\} \in \mathcal{W}_p\}$ for any $(A_1, ..., A_n) \in Domain(F)$. So we show that $\mathcal{W}_p = \mathcal{W}_q$ for any $p, q \in X$. Consider any $p, q \in X$. We show that $\mathcal{W}_p \subseteq \mathcal{W}_q$; the inclusion $\mathcal{W}_q \subseteq \mathcal{W}_p$ holds analogously. We suppose $C \in \mathcal{W}_p$ and show $C \in \mathcal{W}_q$. If $C = N$, then $N \in \mathcal{W}_q$ by claim 2. Now assume $C \neq N$. We have $C \neq \emptyset$; otherwise there would be a profile $(A_1, ..., A_n) \in Domain(F)$ with $p \notin A_i$ for each i and $p \in F(A_1, ..., A_n)$, whence (by the completeness and consistency of each individual and collective judgment set) $\neg p \in A_i$ for each i and $\neg p \notin F(A_1, ..., A_n)$, violating claim 1.

Since C is neither $\emptyset$ nor N and $C \in \mathcal{W}_p$, there is a $(A_1, ..., A_n) \in Domain(F)$ such that some A_i contains p and some A_i contains $\neg p$. So, as each A_i is consistent, p and $\neg p$ are each consistent, i.e. p is contingent. Hence, as X is path-connected, there are $p = p_1, p_2, ..., p_k = q \in X$ with $p_1 \vDash^* p_2$, $p_2 \vDash^* p_3$, ..., $p_{k-1} \vDash^* p_k$. We show by induction that $C \in \mathcal{W}_{p_j}$ for all $j = 1, 2, ..., k$. If $j = 1$ then $C \in C_{p_1}$ by $p_1 = p$. Now let $1 \leq j < k$ and assume $C \in \mathcal{W}_{p_j}$. By $p_j \vDash^* p_{j+1}$, 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 $(A_1, ..., A_n) \in Domain(F)$ 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, $Y \subseteq F(A_1, ..., A_n)$ by claim 1. Since $\{i : p_j \in A_i\} = C \in \mathcal{W}_{p_j}$, we have $p_j \in F(A_1, ..., A_n)$. So $\{p_j\} \cup Y \subseteq F(A_1, ..., A_n)$. Hence, since $\{p_j, \neg p_{j+1}\} \cup Y$ is inconsistent, $\neg p_{j+1} \notin F(A_1, ..., A_n)$, whence $p_{j+1} \in F(A_1, ..., A_n)$. So, as $\{i : p_{j+1} \in A_i\} = C$, we have $C \in \mathcal{W}_{p_{j+1}}$, as desired.

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 $(A_1, ..., A_n) \in Domain(F)$ with $\{i : p \in A_i\} = C$; this holds because X contains a contingent proposition p. Part (2) holds because, for any $(A_1, ..., A_n) \in Domain(F)$, each of the sets $A_1, ..., A_n, F(A_1, ..., A_n)$ 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 $p = p_1, p_2, ..., p_k = \neg p \in X$ and $Y_1^*, Y_2^*, ..., Y_{k-1}^* \subseteq X$ such that

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

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

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

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

We first show that there exists a $t \in \{1, ..., k-1\}$ such that $\{p_t, \neg p_{t+1}\}$ is consistent. Assume for contradiction that each of $\{p_1, \neg p_2\}, ..., \{p_{k-1}, \neg p_k\}$ is inconsistent. So (using L2) each of $\{p_1, \neg p_2\}, \{p_1, p_2, \neg p_3\}, ..., \{p_1, ..., p_{k-1}, \neg p_k\}$ 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 $\{p_1, ..., p_k\}$ 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 $t \in \{1, ..., k-1\}$ such that $\{p_t, \neg p_{t+1}\}$ is consistent, whence $Y_t^* \neq \emptyset$ by (*). 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 p or $\neg p$ (but not both) for each $p \in Y_t^*$ (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 (*),
- 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 (**),
- $(Y \backslash Z_2) \cup Z_2^{\neg} = (Y \backslash \{\neg p_{t+1}\}) \cup \{p_{t+1}\} = \{p_t, p_{t+1}\} \cup Y_t^*$ is consistent by (**),
- $(Y \backslash Z_3) \cup Z_3^{\neg} = \{p_t, \neg p_{t+1}\} \cup (Y_t^* \backslash Z_3) \cup Z_3^{\neg} = \{p_t, \neg p_{t+1}\} \cup B$ 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 $(A_1, ..., A_n)$ 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 $Z_1 \subseteq F(A_1, ..., A_n)$. 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 $Z_2 \subseteq F(A_1, ..., A_n)$. Further, $Z_3 \subseteq F(A_1, ..., A_n)$ as $C_1 \cup C_2 = C^* \in \mathcal{W}$. Finally, $Y \backslash (Z_1 \cup Z_2 \cup Z_3) \subseteq F(A_1, ..., A_n)$ as $N \in \mathcal{W}$ by claim 3. In summary, we have $Y \subseteq F(A_1, ..., A_n)$, violating consistency.

Claim 6. There is a dictator.

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

If the agenda is not path-connected, then there may exist non-dictatorial aggregation rules satisfying all of Theorem 2's conditions; examples of such agendas are not only trivial agendas (containing a single proposition-negation pair or several logically independent such pairs), but also *network agendas*, including the simple network agenda defined above (Dietrich [9]).

By contrast, if we consider *rich* agendas, special cases of path-connected agendas characterized by a very particular logical structure, then we obtain an impossibility result even if we weaken the responsiveness condition in Theorem 2.[12]

Weak Responsiveness. The rule is not constant: 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 a rich agenda, 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.*

Proof. By Theorem 1, under universal domain, non-manipulability is equivalent to the conjunction of independence and monotonicity. Now the result follows immediately from theorems by Pauly and van Hees [26] (when X is atomically closed) and Dietrich [7] (when X is atomic). ■

While the impossibility of Theorem 2 applies to conjunctive (and disjunctive) agendas, the impossibility of Theorem 3 does not, as these agendas are not rich. A non-dictatorial aggregation rule for a conjunctive agenda (with $k \geq 2$) satisfying the conditions of Theorem 3 is given by taking a majority vote on a_1 and always accepting $\neg a_2, \ldots, \neg a_k$, $\neg c$ and $c \leftrightarrow (a_1 \wedge \cdots \wedge a_k)$. This rule is weakly responsive, but not responsive.

[12] An agenda X is *rich* if (i) X contains at least two contingent propositions p and q (with p not equivalent to q or $\neg q$) and (ii) X is atomically closed or atomic. Here X is *atomically closed* if (i) X belongs to standard 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 [26]). 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 and is inconsistent with a member of each pair $q, \neg q \in X$ (Dietrich [7]); intuitively, X is atomic if the disjunction of the atoms of X is a tautology (its negation is inconsistent).

3.5 Escape-routes from 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. For a general agenda, *propositionwise supermajority rules* – requiring a particular supermajority (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 compact logic or a finite agenda, a supermajority rule requiring at least m votes for the acceptance of any proposition guarantees collective consistency if and only if $m > n - n/z$, where z is the size of the largest minimal inconsistent set $Z \subseteq X$ (Dietrich and List [11]; List [20]).

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 complete and consistent individual judgment sets satisfying *unidimensional alignment*, a structure condition similar to single crossing in preference aggregation (List [18]). A profile $(A_1, \ldots, A_n)$ is *unidimensionally aligned* if, for each $p \in X$, there exists a linear ordering Ω on N such that [for all $i \in N_p^+$ and all $j \in N_p^-$, $i \Omega j$] or [for all $i \in N_p^-$ and all $j \in N_p^+$, $i \Omega j$], where $N_p^+ := \{i \in N : p \in A_i\}$ and $N_p^- := \{i \in N : p \notin A_i\}$. For a related result on preference aggregation, see Saporiti and Thomé [29].

4 Strategy-proofness

Non-manipulability is not yet a game-theoretic concept. We now define strategy-proofness in a game-theoretic sense and prove its equivalence to non-manipulability as defined above.

4.1 Preference relations over judgment sets

We interpret a judgment aggregation problem as a game whose players are the n individuals. 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, $A, B \subseteq X$, $A \succsim_i B$ 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.[13] (Individuals need not be able to rank all pairs of judgment sets relative to each other.)

How can an individual's preference relation over different collective judgments sets be motivated? An *epistemically motivated* individual prefers judgment sets that she considers closer to the truth (where she might consider her own judgment set as the truth). A *non-epistemically motivated* individual prefers judgment sets for reasons other than the truth, for example because she might personally benefit from decisions resulting from the collective endorsement of some judgment sets rather than others.

One can make different assumptions on how an individual's preference relation is related to her true judgment set. To formulate such assumptions, we introduce 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. Different specifications of C are appropriate for different groups of individuals and different aggregation problems. Let us now mention some important examples (in an order of increasing 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. $C(A_i) = \{\succsim\text{: } A_i \succsim B \text{ for all judgment sets } B\}$.

Closeness-respecting preferences. We say that a judgment set B is *at least as close* to A_i as another judgment set B^* if, for all propositions $p \in X$, 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 closer to $\{a, b, c \leftrightarrow (a \wedge b), c\}$ than $\{\neg a, \neg b, c \leftrightarrow (a \wedge b), \neg c\}$,[14] 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 $\{a, b, c \leftrightarrow (a \wedge b), c\}$. We also say that a preference relation $\succsim$ *respects closeness* to A_i if, for any two judgment sets B and B^*, if B is at least as close to A_i as B^*, then $B \succsim B^*$. Now, for each A_i, $C(A_i)$ is the set of all preference relations $\succsim$ that respect closeness to A_i, and we write $C = C_X$. One element of $C_X(A_i)$ is the (complete) preference relation induced by the Hamming distance to A_i.[15] Our definition of closeness between judgment sets is related to Schulte's [31] definition of Pareto-minimal theory change.

Y-oriented preferences (for some $Y \subseteq X$). Generalizing our earlier definitions, a judgment set B is *at least as close* to A_i *on* Y as another judgment set B^* if, for

[13] $\succsim_i$ is: *reflexive* if, for any judgment set A, $A \succsim_i A$; *transitive* if, for any judgment sets A, B, C, $A \succsim_i B$ and $B \succsim_i C$ implies $A \succsim_i C$; *complete* if, for any judgment sets A, B, $A \succsim_i B$ or $B \succsim_i A$.

[14] Here "closer than" is the strong component of "at least as close as".

[15] The Hamming distance between two judgment sets B and B^* is the number of propositions $p \in X$ such that B and B^* disagree on p, written $d(B, B^*)$. The preference relation $\succeq$ induced by Hamming distance to A_i is defined, for any judgment sets B, B^*, by $[B \succeq B^*$ if and only if $d(B, A_i) \leq d(B^*, A_i)]$. In the special case of the Arrow agenda, recall that each judgment set represents a preference ordering on the set of options K. Here an individual's preference relation $\succeq_i$ over judgment sets represents her meta-preference over preference orderings. In their work on strategy-proofness of social welfare functions, Bossert and Storcken [4] use the Kemeny distance between preference orderings to obtain such a meta-preference. On distances between preference orderings, see also Baigent [2].

all propositions $p \in Y$, if B^* agrees with A_i on p, then B also agrees with A_i on p; and 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 $B \succsim B^*$. On the assumption of Y-oriented preferences, 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 $C = C_Y$. This captures the case where individuals care only about the propositions in Y. When $Y = X$, Y-oriented preferences coincide with closeness-respecting preferences *simpliciter*. If $Y_1 \subseteq Y_2$, then, for all A_i, $C_{Y_1}(A_i) \subseteq C_{Y_2}(A_i)$. Below we analyse the special cases of "reason-oriented" and "outcome-oriented" preferences.

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 every individual i, profile $(A_1, \ldots, A_n) \in Domain(F)$ 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 $(A_1 \ldots, A_i^*, \ldots, A_n) \in Domain(F)$.

Our definition of strategy-proofness (generalizing List [20]) is similar to the classical one given by Gibbard [15] and Satterthwaite [30]. For certain specifications of C, there are parallels to the approaches in Barberà et al. [1] and Nehring and Puppe [24], but an important disanalogy is the different input of the aggregation rule in our model (each individual submits a single judgment set rather than a preference relation over points in a hypercube).

If the domain of F is a Cartesian product domain (such as the universal domain), then strategy-proofness implies that truthfulness is a weakly dominant strategy for every individual. If the domain is not a product domain, then 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 [29].

As in the case of non-manipulability above, we have defined a family of strategy-proofness conditions, one for each specification of C. If two functions C_1 and C_2 are such that, for each A_i, $C_1(A_i) \subseteq C_2(A_i)$, then strategy-proofness for C_2 implies strategy-proofness for C_1.

4.3 The equivalence of strategy-proofness and non-manipulability

What is the logical relation between non-manipulability as defined above and strategy-proofness?

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

So strategy-proofness for Y-oriented preferences is equivalent to non-manipulability on Y. In particular, strategy-proofness for C_X is equivalent to non-manipulability

simpliciter. We may therefore consider strategy-proofness for C_X as our default condition, equivalent to our default condition of non-manipulability.

Proof. 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 $(A_1, ..., A_n) \in Domain(F)$, such that $F(A_1, ..., A_n)$ disagrees with A_i on p. Let $(A_1, ..., A_i^*, ..., A_n) \in Domain(F)$ be any i-variant. We have to show that $F(A_1, ..., A_i^*, ..., A_n)$ 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, $F(A_1, ..., A_n) \succsim_i F(A_1, ..., A_i^*, ..., A_n)$. Since A_i disagrees with $F(A_1, ..., A_n)$ on p, the definition of $\succsim_i$ implies that A_i still disagrees with $F(A_1, ..., A_i^*, ..., A_n)$ on p.

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

Using Theorem 4, we can now restate Theorems 1 and 2 in terms of strategy-proofness:

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, (ii) and (iii) are equivalent, and each implies (i).

We write $C \supseteq C_X$ to mean that, for each A_i, $C(A_i) \supseteq C_X(A_i)$.

Corollary 2 *Let $C \supseteq C_X$. For a path-connected agenda (e.g. a conjunctive, disjunctive or Arrow agenda), an aggregation rule F satisfies universal domain, collective rationality, responsiveness and strategy-proofness for C if and only if F is a dictatorship of some individual.*

In particular, if the individuals' preferences over judgment sets are unrestricted, top-respecting or closeness-respecting, we obtain an impossibility result. As before, for a rich agenda, responsiveness can be relaxed to weak responsiveness.

5 An application: 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 Y-oriented ($C = C_Y$) for some subset $Y \subsetneq X$, 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 Definitions

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

Outcome-(or conclusion-)oriented preferences. $C = C_{Y_{outcome}}$, where $Y_{outcome} = \{c, \neg c\}$.

Reason-(or premise-)oriented preferences. $C = C_{Y_{reason}}$, where $Y_{reason} = \{a_1, \neg a_1, \ldots, a_k, \neg a_k\}$.

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 material consequences that 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 (such as setting precedents for future decisions). Proponents of "deliberative democracy" often make the motivational assumption of reason-oriented preferences.

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 $Y_{outcome} = \{c, \neg c\}$, whereas A is closer to A_3 than A^* on $Y_{reason} = \{a, \neg a, b, \neg b\}$. 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, premise-based voting is strategy-proof for* $C_{Y_{reason}}$.

We prove this result directly, although it can also be derived from Corollary 1.

Proof. 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 $F(A_1, \ldots, A_n) \succsim_i F(A_1, \ldots, A_i^*, \ldots, A_n)$. ■

This result is interesting from a "deliberative democracy" perspective. If individuals have reason-oriented preferences in deliberative settings (as sometimes assumed by proponents of "deliberative 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, 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* $Z \subseteq X$ *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 $p \in Z$.

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

Proof. 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 $(A_1, ..., A_n) \in Domain(F) = Domain(G)$ and any preference relations $\succsim_1 \in C_{Y_{outcome}}(A_1)$, ..., $\succsim_n \in C_{Y_{outcome}}(A_n)$. Define $(B_1, \ldots, B_n)$ by

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

It can easily be seen that, for each i and any pair of i-variants $(D_1, \ldots, B_i, \ldots, D_n)$, $(D_1, \ldots, B_i^*, \ldots, D_n) \in Domain(F)$, $F(D_1, \ldots, B_i, \ldots, D_n)$ is at least as close to A_i on $Y_{outcome}$ $(= \{c, \neg c\})$ 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 $(A_1, ..., A_n)$ (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 $G(C_1, \ldots, C_n) = G(A_1, ..., A_n)$ agree on each proposition in $Y_{outcome} = \{c, \neg c\}$. ■

Despite the differences between premise- and conclusion-based voting, if individuals have outcome-oriented preferences and act on appropriate weakly dominant strategies, then 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 Summary

This paper is the first investigation of strategic manipulation and strategy-proofness in judgment aggregation. We have introduced a non-manipulability condition for judgment aggregation and characterized the class of non-manipulable aggregation rules. We have then defined a game-theoretic strategy-proofness condition and shown its equivalence to non-manipulability, as defined earlier. Given this equivalence, we have obtained a characterization of strategy-proof aggregation rules. We have also proved an impossibility result which is the judgment aggregation analogue of the classical Gibbard-Satterthwaite theorem. For the general class of path-connected agendas, including conjunctive, disjunctive and Arrow agendas, all strategy-proof aggregation rules satisfying some mild conditions are dictatorial.

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*. Likewise, instead of requiring strategy-proofness for a large set of individual preferences over judgment sets, we may require strategy-proofness only for a restricted set of preferences, for example for "outcome-" or "reason-oriented" preferences. Premise-based voting, for example, is strategy-proof for "reason-oriented" preferences.

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 challenge 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 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.

References

[1] Barberà, S., Massóa, J., Nemeb, A.: Voting under constraints. Journal of Economic Theory 76(2), 298-321 (1997)

[2] Baigent, N.: Preference Proximity and Anonymous Social Choice. Quarterly Journal of Economics 102(1), 161-170 (1987)

[3] Bovens, L., Rabinowicz, W.: Democratic Answers to Complex Questions - An Epistemic Perspective. Synthese, forthcoming (2005)

[4] Bossert, W., Storcken, T.: Strategy-proofness of social welfare functions: the use of the Kemeny distance between preference orderings. Social Choice and Welfare 9, 345-360 (1992)

[5] Brennan, G.: Collective Coherence? International Review of Law and Economics 21(2), 197-211 (2001)

[6] Chapman, B.: Rational Aggregation. Politics, Philosophy and Economics 1, 337-354 (2002)

[7] Dietrich, F.: Judgment Aggregation: (Im)Possibility Theorems. Journal of Economic Theory, forthcoming (2005)

[8] Dietrich, F.: Judgment aggregation in general logics. Working paper, PPM Group, University of Konstanz (2004)

[9] Dietrich, F.: The possibility of judgment aggregation for network agendas. Working paper, PPM Group, University of Konstanz (2005)

[10] Dietrich, F., List, C.: Arrow's theorem in judgment aggregation. Working paper, LSE (2005)

[11] Dietrich, F., List, C.: Judgment aggregation by quota rules. Working paper, LSE (2005)

[12] Dokow, E., Holzman, R.: Aggregation of binary evaluations. Working paper, Technion, Israel (2005)

[13] Ferejohn, J.: Conversability and Collective Intention. Paper presented at the Common Minds Conference, Australian National University, 24-25 July 2003 (2003)

[14] Gärdenfors, P.: An Arrow-like theorem for voting with logical consequences. Economics and Philosophy, forthcoming (2005)

[15] Gibbard, A.: Manipulation of voting schemes: a general result. Econometrica 41, 587-601 (1973)

[16] van Hees, M.: The limits of epistemic democracy. Working paper, University of Groningen (2004)

[17] Kornhauser, L. A., Sager, L. G.: Unpacking the Court. Yale Law Journal 96(1), 82-117 (1986)

[18] List, C.: A Possibility Theorem on Aggregation over Multiple Interconnected Propositions. Mathematical Social Sciences 45(1), 1-13 (2003)

[19] List, C.: The Probability of Inconsistencies in Complex Collective Decisions. Social Choice and Welfare 24(1), 3-32 (2005)

[20] List, C.: A Model of Path Dependence in Decisions over Multiple Propositions. American Political Science Review 98(3), 495-513 (2004)

[21] List, C., Pettit, P.: Aggregating Sets of Judgments: An Impossibility Result. Economics and Philosophy 18, 89-110 (2002)

[22] List, C., Pettit, P.: Aggregating Sets of Judgments: Two Impossibility Results Compared. Synthese 140(1-2), 207-235 (2004)

[23] Nehring, K., Puppe, C.: Consistent Judgement Aggregation: A Characterization. Working paper, University of Karlsruhe (2005)

[24] Nehring, K., Puppe, C.: The Structure of Strategy-Proof Social Choice: General Characterization and Possibility Results on Median Spaces. Working paper, University of Karlsruhe (2004)

[25] Osherson, D., Vardi, M.: Aggregating Disparate Estimates of Chance. Games and Economic Behavior, forthcoming (2005)

[26] Pauly, M., van Hees, M. Logical Constraints on Judgment Aggregation. Journal of Philosophical Logic, forthcoming (2005)

[27] Pettit, P.: Deliberative Democracy and the Discursive Dilemma. Philosophical Issues 11, 268-299 (2001)

[28] Pigozzi, G.: Collective decision-making without paradoxes: An argument-based account. Working paper, King's College, London (2004)

[29] Saporiti, A., Thomé, F.: Strategy-proofness and single-crossing. Working, Queen Mary, University of London (2005)

[30] Satterthwaite, M.: Strategyproofness and Arrow's conditions: existence and correspondences for voting procedures and social welfare functions. Journal of Economic Theory 10, 187-217 (1975)

[31] Schulte, O.: Minimal belief change, Pareto-optimality and logical consequence, Economic Theory 19(1), 105-144 (2005)

[32] Wilson, R.: On the Theory of Aggregation. Journal of Economic Theory 10, 89-99 (1975)

A Appendix

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

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

where a proposition is of type 1 if it is a possibly negated premise ($a_1, \neg a_1, ..., a_k$, $\neg a_k$), 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 (*) holds, then, for any $p, q \in X$ of the *same* type, taking any $s \in X$ of a different type, there is by (*) 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 $p \vDash^* q$ if and only if $\neg q \vDash^* \neg p$ (use both times the same Y), claim (*) is equivalent to

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

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

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

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

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