MANIPULATING VOTING PROCEDURES

ALLAN FELDMAN*

A voting procedure can be manipulated *if, by misrepresenting his preferences, some individual can secure an outcome which he prefers to the outcome he gets when he is honest.*

This is an expository paper on the theory of voting manipulation. Section I is an historical sketch of the contributions of Condorcet, de Borda, Arrow, and others. Section II provides a set of examples of manipulation: of plurality voting, of majority voting, of exhaustive voting, of the single transferable vote procedure, and of approval voting. It also contains an example of a nonmanipulable random voting scheme. Section III provides a simple proof of the Gibbard-Satterthwaite manipulation theorem.

I. INTRODUCTION

The French Enlightment left Western Civilization with, among other things, the first systematic analyses of the properties of elections. These analyses grew out of the then blossoming interest in democratic institutions and the democratic or egalitarian state. However, French political philosophers, particularly Rousseau, de Borda, and Condorcet, may have raised as many questions as they answered about the nature of elections as expressions of the general will. The questions they raised are with us still, because, unfortunately, elections are logically imperfect. The purpose of this paper is to discuss in a relatively nontechnical way the nature of logical imperfections of elections.

Let us begin near the beginning, with the Marquis de Condorcet (1743-1794). (For a detailed exposition of Condorcet's theory, as well as most other historically important voting theories, see Duncan Black (1958).) In the *Essai sur l'Application de l'Analyse à la Probabilité des Decisions Rendues à la Pluralité des Voix* (Paris, 1785) he set out to solve the following problem in probability theory: A jury is to decide, by voting, between two alternatives A and B (which might be, say, the guilt or innocence of a defendant). One alternative is right, and the other is wrong. The members of the jury, however, are imperfect. When they vote they may err. Given particular probabilities of each member's voting correctly or incorrectly, what is the probability that the jury's

*Brown University. Acknowledgements: Peter Fishburn, Allan Gibbard, Elisha Pazner and David Schmeidler made helpful comments on an earlier draft. Mark Grimm assisted with some of the examples.

Economic Inquiry
Vol. XVII, July 1979

decision is right? This is not difficult to calculate; naturally it depends on the numbers of jury members voting for A and B. However, the analysis becomes more complicated when there are three (or more) alternatives, say, A, B and C. Again, one alternative is right, and the others are wrong. If votes are taken by the jury between pairs of alternatives, what is the probability that the jury's decision is correct? The question is meaningful when the voting results are consistent, for example, if A gets more votes than B in a contest between the two, if B gets more votes than C in a contest between those two, and if A gets more votes than C in a contest between those two. In this case the jury's decision is A, and the probability that this decision is right can be calculated in a straightforward manner.

However, there are cases in which the voting between pairs of alternatives *cycles*. In such cases there do exist straightforward ways to calculate the probabilities of correctness of each of the alternatives, but these might all be distressingly small. In fact, the cycling cases prompted Condorcet to switch from obvious probability calculations to a somewhat an *ad hoc* judgment, the essence of which is this: if an alternative X is decisively defeated by another alternative Y, X cannot be the right alternative. This judgment is not terribly satisfactory and will not play an important role in what follows. However, the cycling or "paradoxical" cases of Condorcet still provide the simplest examples of voting imperfections, so we turn to one now:

Assume there are seven voters of three types, and three alternatives, A, B and C. There are three type one voters, and each type one voter prefers A to B to C. There are two type two voters, each of whom prefers B to C to A. There are two type three voters, each of whom prefers C to A to B. These assumptions are illustrated below:

	Type 1	Type 2	Type 3
Order	A	B	C
of	B	C	A
Preference	C	A	B
	(3 voters)	(2 voters)	(2 voters)

In a vote between A and B, A wins, five votes (from type 1 and 3 voters) to two (from type 2 voters). In a vote between B and C, B gets five votes (from type 1 and 2 voters) and C gets two (from type 3 voters). In a vote between A and C, however, C gets four votes (from type 2 and 3 voters) and A gets three (from type 1 voters). Thus there is a cycle in the voting results: A defeats B, B defeats C, but C defeats A. In terms of the jury problem, a cyclical structure like this might result in unacceptably small (correctly calculated) probabilities of rightness for *every* alternative.

Condorcet's *ad hoc* judgment would be to discard B and C, since each of these is defeated five to two, while A is (at worst) defeated four to three.

But if this judgment strikes us as unacceptably arbitrary, or if the probability analysis is fundamentally unacceptable because a "correct" choice is unknown, unknowable and/or non-existent, then *majority voting between pairs fails as a method for making social choices.*

There are, however, other methods. Jean-Charles de Borda (1733-1799) published his *Memoire sur les Élections au Scrutin* (see De Grazia (1953)) four years before Condorcet's *Essai*. In it de Borda developed a "method of marks." Each elector ranks the alternatives according to his order of preference (ignoring the possibility of ties). If there are k alternatives, an elector's first choice is assigned k points, his second k-1 points, and so on down to his last choice, which is assigned one point. The total vote for an alternative is the sum of the points assigned it by the various electors, and the winner (barring ties) is the alternative with the highest sum.

For the example above, the de Borda votes are:

$$3x3 + 2x1 + 2x2 = 15 \text{ for A},$$
$$3x2 + 2x3 + 2x1 = 14 \text{ for B},$$
$$3x1 + 2x2 + 2x3 = 13 \text{ for C},$$

so A wins. Clearly, de Borda's method escapes the cycling possibility of pairwise majority voting, since the vote totals of the alternatives cannot cycle. However, the method *is* problematic.

The problem we focus on, which is relayed secondhand by Black (1958, p. 182) and noted by Satterthwaite (1975), is that of deliberate misrepresentation of their preferences by the electors. De Borda is quoted as saying "my scheme is only intended for honest men." To illustrate just how appropriate this remark is, we modify the example by adding two alternatives, D and E, and assuming the following structure of preferences:

Type 1	Type 2	Type 3
A	B	C
B	C	A
C	A	B
D	D	D
E	E	E
(3 voters)	(2 voters)	(2 voters) .

Now the de Borda counts are:

$$3x5 + 2x3 + 2x4 = 29 \text{ for A,}$$
$$3x4 + 2x5 + 2x3 = 28 \text{ for B,}$$
$$3x3 + 2x4 + 2x5 = 27 \text{ for C,}$$
$$3x2 + 2x2 + 2x2 = 14 \text{ for D,}$$
$$3x1 + 2x1 + 2x1 = 7 \text{ for E,}$$

and A wins again. However, if one of the type 2 electors had falsely declared

B
C
D
E
A

as his preference ordering, the de Borda counts would have been 28 for B, 27 for C and 27 for A, and B would have won. This elector would have been better off than when he voted honestly: *the method provides a temptation for misrepresentation of preferences, or "strategic" voting.*

The possibility of manipulating the result of an election through the misrepresentation of preferences was not seriously considered by either de Borda or Condorcet. The Rev. C. L. Dodgson gave it some passing consideration almost a century later, especially in Chapter I, Section 5 of "A Discussion of the Various Methods of Procedure in Conducting Elections" (1873), reprinted in Black (1958). Dodgson's "method of marks" procedure (which differs from de Borda's) works as follows:

> "a certain number of marks is fixed, which each elector shall have at his disposal; he may assign them all to one candidate, or divide them among several candidates, in proportion of their eligibility; and the candidate who gets the greatest total of marks is the winner."

Dodgson writes the method would

> "be absolutely perfect, if only each elector wished to do all in his power to secure *that candidate who should be the most generally acceptable*"

(his italics); however,

> "we are not sufficiently unselfish and public-spirited to give any hope of this result being attained."

Each elector would attempt to manipulate the results by assigning all of his votes to his own favorite candidate. In fact, this is not quite correct, as a voter might assign all his marks to his second choice if he thought his first could not win, but the method is obviously easily manipulable. Although Dodgson was aware of the potential of election manipulation through voter misrepresentation of preferences, there is no reason to believe he thought the problem inescapable, and the question lay dormant for most of yet another century.

The modern surge in interest in properties of voting procedures began in the late 1940's and early 1950's with two results, one modest and one profound. In 1948 Duncan Black "solved" the cyclical voting paradox by deriving conditions under which pairwise majority voting cannot cycle. However, this interesting result had limited ultimate bearing on the question of election manipulation, or strategic voting. (Limited bearing is not no bearing; see Blin and Satterthwaite (1976).) The profound result, which ultimately did bear heavily on the question of manipulability of elections, was Kenneth Arrow's 1951 impossibility theorem for preference aggregation procedures (Arrow (1963)).

At least in its early incarnations, Arrow's theorem was not technically about voting procedures. Voting procedures generate single alternatives (winners) from among sets of alternatives; Arrow's theorem is on its surface about procedures which generate orderings of the entire set of alternatives. To be more precise, we need to introduce some terminology. It is assumed that each person in society has an ordering, or ranking, of the set of alternatives. A specification of all people's orderings is called a *preference profile*. For example, in the Condorcet voting paradox example,

Type 1	Type 2	Type 3
A	B	C
B	C	A
C	A	B
(3 voters)	(2 voters)	(2 voters)

is a preference profile.

What we have called voting procedures are rules for transforming preference profiles into winners, or mappings from the set of possible preference profiles into the set of alternatives. For each preference profile the mapping produces a single winning alternative. Technically such a mapping is called a *social decision function*, or SDF. An SDF takes a preference profile, digests it, and produces a winning alternative. (A more generally defined SDF maps preference profiles into *sets* of winning alternatives, rather than single winners. See, e.g., Sen (1970) and (1977). This paper, however, is only about single-winner SDF's.) In contrast,

the rules Arrow first studied, which are now called *social welfare functions*, or SWF's, are rules for transforming preference profiles into social preference orderings or rankings. An SWF takes a preference profile, digests the list, and produces a social ordering. Obviously, there are procedures which can be viewed as either SWF's or as SDF's; de Borda voting is one.

The question Arrow asked is whether or not there exists a "satisfactory" SWF. The answer naturally depends on what is incorporated into "satisfactory," or what properties one wants the SWF to have. Arrow's criteria, which have been somewhat modified and refined over the years, were basically as follows:

(1) The SWF must produce a social *ordering*, that is a complete, reflexive and transitive social preference relation. This requirement excludes majority voting between pairs, which can give rise to (intransitive) cycles.

(2) The SWF must always work, no matter what (finite) set of alternatives and preference profiles are given to it.

(3) The SWF must respond positively to individual preferences. Loosely speaking, if x is socially preferred to y at the start, and x gains support, x must remain preferred to y at the end.

(4) The SWF cannot be imposed. That is, if all individuals prefer x to y, the SWF must yield a social ordering which ranks x above y.

(5) The SWF must show "independence of irrelevant alternatives." This means that the social ranking of x vis-à-vis y must depend *only* on individual rankings of x vis-à-vis y, not on the strengths of feelings, not on rankings of x vs. z or y vs. w or z vs. w, or any other such "irrelevancy." It is not difficult to see that de Borda's method violates independence, as does Dodgson's method of marks.

(6) There must be no dictator, no single person whose individual preference ordering always defines the social preference ordering.

In his remarkable theorem, Arrow showed that *no "satisfactory" SWF exists.*

This negative answer might seem to shadow the search for a perfect voting procedure, or SDF. If there is no satisfactory way to aggregate individual preferences into a social preference ordering, perhaps there is no satisfactory way to aggregate individual preferences into an election winner. But the connection is really not so obvious. In its original form, Arrow's theorem is about generating social preference relations. Although the theorem can be, and has been, translated into a collection of theorems about what we would call generally-defined SDF's (see, e.g., Sen (1977), pp. 71-75), there are hurdles in the translation. For example, it is not immediately clear that an SDF, a procedure which merely generates winners, can be adapted in some way that accommodates

Arrow's requirements (1), (3), (4) or (5), or that those requirements can be adapted in a way which makes sense for SDF's. Moreover, it seems intuitively reasonable that generating single alternatives (winners) ought to be less of a *strain* on a decision procedure than generating whole lists, or orderings, of the alternatives, so if the latter is impossible the former might not be.

In fact, Arrow's theorem profoundly affected the search for an ideal SDF in three ways. First, its negative conclusion (there is *no* satisfactory SWF) generated an enormous intellectual storm, prompting some to try to show what Arrow did wrong, and others to show how the results could be generalized. Some of the storm's electricity leaked to the analysis of SDF's. Second, Arrow's theorem was actually used as a tool by some in proving the impossibility theorem concerning SDF's which will occupy us below (Gibbard (1973); and Schmeidler and Sonnenschein (forthcoming), proof I). Third, and perhaps most important, Arrow's theorem suggests two questions about SDF's which are analogous to the questions Arrow raised about SWF's: What is a "satisfactory" SDF? Does a "satisfactory" SDF exist?

What then is a "satisfactory" SDF? The remarks about manipulation of the de Borda and Dodgson rules suggests one possible criterion: An SDF ought to be immune to manipulation through the misrepresentation of preferences. Moreover, common sense suggests that a "satisfactory" SDF shouldn't be a dictatorship. These two requirements may not seem like much in light of the six occasionally complex requirements Arrow imposes on SWF's. In fact, however, several authors conjectured around 1960 (Dummett and Farquharson (1961) and Vickrey (1960)) that any nondictatorial voting procedure is manipulable, and finally in the 1970's Allan Gibbard (1973) and Mark Allen Satterthwaite (1973, 1975) independently proved this is the case. *If a satisfactory social decision function is one which is always immune to manipulation and which is nondictatorial, there is no satisfactory social decision function.*

This paper provides, in Section II, a number of detailed examples of manipulation, and in Section III, a nontechnical discussion of the Gibbard-Satterthwaite result and a heuristic version of the proof of their theorem.

II. MORE EXAMPLES OF MANIPULABILITY AND NON-MANIPULABILITY

In this section several voting procedures are analyzed to discover whether or not they are manipulable. The first four procedures fit the technical definition for SDF's: each takes every voter's preference ordering (the preference profile) as inputs, and produces a single, certain winner as its output. The fifth is not a SDF; it takes every voter's set of acceptable or approved alternatives as inputs, and produces a certain winner as its output. (A voter's acknowledgement that "X, Y and Z are

acceptable, whereas W, U, V, . . . are unacceptable" is not quite the same thing as his declaring $\begin{matrix} Y \\ Z \\ X \\ U \\ W \\ \cdot \\ \cdot \\ \cdot \end{matrix}$ as his preference ordering.)

The sixth procedure is a lottery type mechanism. It takes every voter's preference ordering as its inputs, and produces *odds* for the alternatives as an output. The actual winner is then chosen randomly. Now to the examples:

The first procedure is the *plurality voting* rule. There are many candidates or alternatives. Each elector casts one vote for one candidate. The candidate with the highest total wins.

This is a common practice, typically used in U.S. party primary and general elections, for example. It also is common that a voter would like to see a candidate with an extreme or "pure" position win, but does not vote for that candidate because to do so would be tantamount to "throwing his vote away." To make one's vote *count*, one votes for a candidate who has a good chance of winning. However, the reluctance to throw one's vote away implies, in our terms, the desire to manipulate. For example, suppose there are three types of voters with the following preferences:

Type 1	Type 2	Type 3
A	B	C
B	C	B
C	A	A
(10 voters)	(9 voters)	(2 voters) .

In a sincere election, type 3 voters cast their votes for C, but A wins the plurality. If type 3 voters anticipate this result, they can vote for B instead; that is, they can vote as if their preferences were $\begin{matrix} B \\ C \\ A \end{matrix}$, and by so doing guarantee that B, whom they prefer to A, is elected. Naturally, type 3 voters would be apt to deny that they are "manipulating" anything; they would say that they are simply not wasting their votes.

Let me note at this point that two voters are manipulating here. Manipulation by a group rather than a single individual is technically

called *coalitional* manipulation, and this and several subsequent examples involve coalitional manipulation. In all these examples of coalitional manipulation, however, it is possible to make modifications to transform them into cases of manipulation by individuals. Unfortunately, the modified examples are slightly inelegant, as they involve tie votes and the resolution of ties by chairmen. (For example, in the case above, if there were 10 type 1 voters and 10 type 2 voters, and if one of the type 1 voters were chairman, then a sincere election would entail 10 votes for A, 10 for B, and 2 for C, and the chairman would break the A-B tie in favor of A. One type 3 voter could then manipulate the election by casting his vote for B, in which case A would get 10, B would get 11, and B would win.) The Gibbard-Satterthwaite theorem of Section III will establish that it is impossible to devise a nondictatorial SDF which is immune to manipulation by *individuals*, and it clearly follows that it is impossible to devise a nondictatorial SDF which is immune to manipulation by coalitions.

The second procedure to look at is *majority voting*, modified by the introduction of an agenda. Because of the possibility of cycling, majority voting between pairs may not give an unambiguous winner, unless the pairwise comparisons are restricted through the use of agendas or other devices. With the preference profile of the voting cycling example, that is,

Type 1	Type 2	Type 3
A	B	C
B	C	A
C	A	B
(3 voters)	(2 voters)	(2 voters)

it has already observed that sincere pairwise majority voting produces a cycle: A defeats B; B defeats C; but C defeats A. Now suppose that A is the "status quo," while B is a motion to change the status quo and C is an amended version of that motion. A typical committee practice (called Procedure α in Black (1958)) is to hold a vote between B and C (the motion and the amended version), and place the winner of that vote against A (the status quo). If voters are sincere, Procedure α produces B on the first round (the amendment is defeated) and A on the second (the bill is defeated).

But under these circumstances, type 2 voters could misrepresent their preferences as $\begin{matrix} C \\ B \\ A \end{matrix}$. If they did, C would win the first round (the amendment would pass) and then C would defeat A (the amended bill would be adopted). This outcome would be preferred by type 2 voters to A, so they

could manipulate the procedure to their benefit.

A second committee practice (called Procedure β in Black (1958)) pits each motion against the status quo, and then selects from among those which defeat the status quo one which defeats the others. In our example, C defeats the status quo while B does not, so C is adopted, provided the voters vote sincerely. But under Procedure β, type 1 voters have an opportunity to gain by misrepresentation. If they vote as if their preferences were $\begin{matrix} B \\ A \\ C \end{matrix}$, both B and C would defeat the status quo in the first round. In the second, B would defeat C, and type 1 voters would have manipulated the choice of B, which they prefer, over C.

Now we turn to somewhat more complex and less often used election rules. The third procedure we consider is the method of *exhaustive voting*, which works in stages. In stage 1, each elector casts a vote for his *least* preferred candidate. The candidate with the largest number of (no-confidence) votes is eliminated from the list. In stage 2, each elector votes for his least-preferred candidate, from the list of remaining candidates. The candidate with the largest number of (no-confidence) votes is again eliminated. The process continues until only one candidate remains, and the one remaining candidate (barring ties) is the winner. For example, suppose the preferences of the electors are as follows:

Type 1	Type 2	Type 3	Type 4	Type 5
A	B	D	C	D
C	C	C	B	C
B	A	B	D	A
D	D	A	A	B
(10 voters)	(7 voters)	(5 voters)	(3 voters)	(4 voters)

The voting goes this way: In stage 1, D is eliminated. In stage 2, A is eliminated. In stage 3, B is eliminated. Therefore, C wins the election when everyone votes sincerely.

The voting could be manipulated by type 1 electors. If they voted as if their preferences were $\begin{matrix} A \\ D \\ B \\ C \end{matrix}$, then in stage 1, C would be eliminated, in stage 2, B would be eliminated, and in stage 3, D would be eliminated. A would be left as the social choice, making type 1 voters better off than they are when they are honest.

The fourth procedure to consider is the method of the *single transferable vote*. This is another staged procedure, in which, at each stage,

voters cast votes for their *most* preferred candidates. In stage 1, each voter casts one vote for his favorite. Then the candidate with the fewest votes is eliminated. In stage 2, each elector casts a single vote for his favorite among the remaining candidates. The candidate with the fewest votes is eliminated. The process continues until one candidate remains; the last remaining candidate is the winner.

With the preferences above, the process works as follows: In stage 1, C, with 3 votes, is eliminated. In stage 2, D, with 9 votes, is eliminated. In stage 3, A, with 14 votes, is eliminated, and B is the ultimate winner. However, the procedure could be manipulated by the type 5 voters. If they voted as if their preferences were

C

D

A

B,

in stage 1, D would be eliminated. In stage 2, B would be eliminated. In stage 3, A would be eliminated, and C would be the winner. And type 5 voters prefer C to B.

The fifth procedure is *approval voting*. (For a detailed analysis, the interested reader should see Brams and Fishburn (1977).) Approval voting works this way: each voter is faced with m candidates, say A, B, C, D, ... He may cast 0, 1, 2, . . . , or even m votes, by assigning a single vote to each candidate he *approves*, and none to each candidate he *disapproves*. The candidate with the highest total wins. For example, consider a simple voting paradox case with three electors. Assume that person 1 is the chairman: if there are ties for first place, he breaks them.

1 (Chairman)	2	3
A	B	C
B	C	A
C	A	B.

Each elector may cast 0, 1, 2, or 3 votes. Casting votes for none or all of the candidates are equally foolish options, which may safely be ignored. Elector 1 might cast one vote for A, one vote for B, and none for C, or one vote for B and none for A or C; and so on. The first two voting strategies involve voicing approval for the top one third or top two thirds of his list by person 1, whereas the third strategy involves his approving B but *not* approving A, whom he really prefers to B. The

first strategies are, therefore, analogous to SDF electors declaring true preferences, and the last is analogous to SDF electors declaring false preferences. Consequently the first two strategies are called *sincere* by Brams and Fishburn. Person 2's sincere strategies are one vote for B and none for C and A; and one vote for B, one vote for C, and none for A. Person 3's sincere strategies are one for C, none for A or B; and one for C, one for A, and none for B.

The question asked about each of the first four procedures was this: Can an example be constructed in which it is advantageous for some person(s) to falsely represent his (their) preferences, *given what the other people are doing?* Let us ask the analogous question for approval voting: Can an example be constructed in which it is advantageous for some person(s) to vote insincerely, given what the other people are doing? The answer is obviously yes. Using the above preferences, suppose each voter casts one vote for his favorite. The results are: one for A, one for B, and one for C. The chairman (person 1) breaks the tie in favor of A. If 2 voted insincerely here, by casting one vote for C and none for B or A, the results would be: one for A, and two for C. So C would win, and two prefers C to A.

In this sense, approval voting *is* manipulable. On the other hand, 2 could also adopt a sincere strategy to secure the desired outcome: if he were to cast one vote for B, one for C and none for A, the results would be: one for A, two for C, one for B, so C would *again* win. In fact, the following proposition is rather obvious: Suppose all voters but i have declared their strategies. Then there exists a *sincere* strategy that will produce the best outcome for i possible, given the strategies of the others. In this sense, approval voting is *not* manipulable.

(Brams and Fishburn discuss manipulability in yet another sense. Given that a voter does not know what the other voters' strategies are, might he vote insincerely in order to hedge, or minimize the risk of an especially bad election outcome? In general, when there are four or more alternatives, the answer is yes.)

Our sixth election procedure is a random one. Given a preference profile, an SDF chooses a single, certain winner. It is a deterministic rule. If a SDF decides twice with the same preference profile, it will produce the same winner both times. Suppose determinism is abandoned. What can be said of a lottery-type social decision mechanism?

Such mechanisms are called *mixed decision schemes*, and the simplest MDS, mentioned, for example, in Gibbard (1973), is a probabilist version of plurality voting: Each voter casts one vote (for his favorite candidate, if he is sincere). Let p_j = the fraction of the vote received by alternative j. Then the winner is *randomly drawn*, with the probability that j wins equal to p_j.

As an example, let the preference profile again be:

Type 1	Type 2	Type 3
A	B	C
B	C	B
C	A	A
((3 voters)	(2 voters)	(2 voters).

Now $p_A = 3/7$, $p_B = 2/7$, and $p_C = 2/7$. A random drawing is performed, in which A's probability of winning is 3/7, B's is 2/7 and C's is 2/7.

That the randomized plurality scheme is immune to manipulation by individuals can be seen as follows: Each voter attempts to maximize an expected utility function

$$EU = p_A U(A) + p_B U(B) + \ldots,$$

where $U(A)$ is the utility to the voter of outcome A, and so on. When the voter casts his vote he affects EU by increasing the relative size of one of the probabilities, and it is clear that EU is increased most when the voter casts his vote for the outcome X for which $U(X)$ is largest. But this implies sincere voting; $U(X)$ is largest for the X the voter likes best.

There is then at least one MDS which is "satisfactory," in the sense that it precludes manipulation. But an MDS won't do, for *randomness per se might be objectionable*. Do we want an election procedure which produces an outcome by a roll of dice? Often not. (Peter Fishburn (1976) has a rather nice discussion, with examples, of unobjectionable social choice lotteries; e.g., the following verse is from the Book of Proverbs in the Bible: "The lot puts an end to disputes and decides between powerful contenders" (18:18). Also see Barberá (1977) and Gibbard (1977) for technical approaches).

Moreover, the particular MDS discussed above has some other problems: It assigns the largest likelihood of choice to A, even though in a pairwise election A would be defeated either by B or by C, and in a de Borda election A would be defeated by B.

This completes our casual survey of some manipulable (and non-manipulable) voting procedures. The first four procedures, all genuine SDF's, are all occasionally liable to manipulation through misrepresentation of preferences. The fifth procedure, approval voting, might or might not be liable to manipulation, depending on the definition of manipulation one adopts. But this procedure is not a genuine SDF. The sixth procedure cannot be manipulated by individuals, but it is a random procedure, and therefore, not a true SDF.

The purpose of this section has been twofold: First, to flesh out the idea of manipulating voting procedures, and second, to show that well-known (and not-so-well-known) SDF's are at least occasionally liable

to manipulation through misrepresentation of preferences. Now it is possible to turn to the theorem that explains why: One can construct manipulation examples for every sensible nondictatorial SDF, well-known, not-so-well-known, or not-yet-invented, because *all such SDF's must be manipulable.*

III. A SIMPLE VERSION OF THE MANIPULATION THEOREM

In this section I will use a very primitive and simple model of society, in which there are only two people and three alternatives (A, B, C). I also suppose that no individual is ever indifferent between two alternatives. (This is essentially the model used by Arrow in a preliminary argument (1963, pp. 48-51), and by me in a simplified proof of Arrow's theorem (Feldman (1974)). An approach similar to what follows is also taken by Schmeidler and Sonnenschein (forthcoming, proof II)).

Person 1 might have any of the following rank orderings of the alternatives:

A	A	B	B	C	C
B	C	A	C	A	B
C	B	C	A	B	A.

The same is true of person 2. Since each of the two can have six preference orderings, there are 6 x 6 = 36 *preference profiles* possible in this society. They are illustrated in Figure 1.

Each cell in this figure shows a preference profile. For example, the cell in the first row, second column,

1	2
A	A
B	C
C	B

,

has person 1 preferring A to B and B to C, and person 2 preferring A to C and C to B.

A social decision function for this little society is a rule which takes every cell of Figure 1, or every preference profile, and transforms it into a winner, or a social choice. For each of the 36 preference profiles in Figure 1, there are three possible social choices. Therefore, the number of conceivable SDF's is 3^{36}, or (approximately) 1.5×10^{17}, or a *hundred and fifty thousand trillion*. Any one of these can be represented by another 6 x 6 matrix, whose entries are the winners (or social choices) corresponding to the preference profiles of Figure 1.

FIGURE 1

Rank Order \ Individuals	1	2	1	2	1	2	1	2	1	2	1	2
1st	A	A	A	A	A	B	A	B	A	C	A	C
2nd	B	B	B	C	B	A	B	C	B	A	B	B
3rd	C	C	C	B	C	C	C	A	C	B	C	A
1st	A	A	A	A	A	B	A	B	A	C	A	C
2nd	C	B	C	C	C	A	C	C	C	A	C	B
3rd	B	C	B	B	B	C	B	A	B	B	B	A
1st	B	A	B	A	B	B	B	B	B	C	B	C
2nd	A	B	A	C	A	A	A	C	A	A	A	B
3rd	C	C	C	B	C	C	C	A	C	B	C	A
1st	B	A	B	A	B	B	B	B	B	C	B	C
2nd	C	B	C	C	C	A	C	C	C	A	C	B
3rd	A	C	A	B	A	C	A	A	A	B	A	A
1st	C	A	C	A	C	B	C	B	C	C	C	C
2nd	A	B	A	C	A	A	A	C	A	A	A	B
3rd	B	C	B	B	B	C	B	A	B	B	B	A
1st	C	A	C	A	C	B	C	B	C	C	C	C
2nd	B	B	B	C	B	A	B	C	B	A	B	B
3rd	A	C	A	B	A	C	A	A	A	B	A	A

Figure 2 represents *one* such SDF:

FIGURE 2

Social Choices

A	A	A	A	A	A
A	A	A	A	A	A
B	B	B	B	B	B
B	B	B	B	B	B
C	C	C	C	C	C
C	C	C	C	C	C

In each cell of Figure 2 is the social choice or social decision derived from the preference profile of the corresponding cell of Figure 1. For example, if the preference profile is

1	2
A	A
B	C
C	B

,

(the first row, second column cell of Figure 1), then the social decision, the winner, is alternative A.

Now Figure 2 represents a very special SDF, for each choice in it is person 1's most preferred alternative! This is a *dictatorial* SDF; it makes 1 a dictator. There is, of course, one other dictatorial SDF; it would be represented by the transpose of the Figure 2 matrix, and it would make 2 a dictator.

The property of SDF's that is of interest here is *manipulability*. How can manipulation be represented in terms of these figures?

Suppose one knows some of the social choices for the preference profiles of row 1 of Figure 1:

Social Choices

A	A	B	C	?	?

(This is part of one possible SDF.) That is, for the preference profiles of row 1, column 1, one knows that A wins; for the preference profile of column 2 A wins; for the preference profile of column 3 B wins; for the preference profile of column 4 C wins, and nothing more is known. If this is the case, person 2 has an opportunity to profitably misrepresent his preferences. Suppose his real preferences are $\begin{matrix}B\\C\\A\end{matrix}$ (column 4 in Figure 1). If he reports this honestly (and 1 is also honest), the SDF outcome is C. However, if he (falsely) claims his preferences are $\begin{matrix}B\\A\\C\end{matrix}$, the SDF outcome is B, which he (truly) prefers to C. In short, person 2 can profitably manipulate the SDF when the preference profile is

1	2
A	B
B	C
C	A

.

If there is *any* opportunity for 1 (or 2) to secure a preferred outcome by misrepresenting his preferences, the SDF is said to be *manipulable*. If it is *never* possible for 1 or 2 to secure a preferred outcome by misrepresentation, the SDF is *nonmanipulable*, or *cheatproof*.

The SDF partly illustrated above is evidently manipulable. What about the dictatorial SDF of Figure 2? Clearly 2 cannot manipulate it since his preferences never affect the outcome. Misrepresenting them must be useless. Nor can 1 manipulate it, since he always gets his (true) first choice. He can never secure a preferred outcome by lying. Dictatorship is, therefore, nonmanipulable.

What of an SDF that is wholly unresponsive to individual preferences? For example, what if an SDF chooses A as the outcome for every preference profile? It obviously would be nonmanipulable. But such SDF's are clearly uninteresting, and we will ignore them by confining our attention to nondegenerate SDF's: a social decision function is *nondegenerate* if each of the three outcomes is chosen for at least one preference profile.

Since there are 150 thousand trillion possible SDF's conceivable in this simple model, it is obviously impossible to systematically examine all of them to discover which, if any, are nonmanipulable. Nonetheless, an unambiguous result is possible, a profound, inescapable, "impossibility" theorem:

Theorem (Gibbard and Satterthwaite): There is no nondegenerate, nondictatorial cheatproof social decision function.

Proof: The proof of the theorem will rely on an intuitively appealing proposition which is proved in the appendix:

Proposition: For a nondegenerate cheatproof SDF, if for some X both persons prefer X to Y, Y cannot be the social choice.

The proposition implies that any nondegenerate nonmanipulable SDF must be consistent with the following figure:

FIGURE 3

Social Choices

A	A	not C	not C	not B	
A	A	not C		not B	not B
not C	not C	B	B		not A
not C		B	B	not A	not A
not B	not B		not A	C	C
	not B	not A	not A	C	C

For example, for the first row, second column preference profile of Figure 1, that is,

1	2
A	A
B	C
C	B

,

A is preferred by both persons to B and C. By the proposition neither B nor C can be the social choice. Therefore, the social choice must be A. Again, for the first row, third column preference profile of Figure 1, that is,

1	2
A	B
B	A
C	C

,

A is preferred by both persons to C, so by the proposition C cannot be the social choice.

Now let us focus on the entire first row of Figure 1:

1	2	1	2	1	2	1	2	1	2	1	2
A	A	A	A	A	B	A	B	A	C	A	C
B	B	B	C	B	A	B	C	B	A	B	B
C	C	C	B	C	C	C	A	C	B	C	A

Thus far, this much is known about the corresponding social choices:

Social Choices

A	A	not C	not C	not B	

To get the machinery cranking, an assumption must be made: Suppose that the social choice for the third column cell (which cannot be C) is A.

It follows that:

Social Choices

A	A	A	not C	not B	

Now if the social choice in column 4, 5 or 6 were B, person 2 would have an opportunity to manipulate in column 3. That is, he could force the choice of B instead of A, when his real preferences are $\begin{matrix} B \\ A \\ C \end{matrix}$, by pretending his preferences are as in 4, 5, or 6. Therefore, for any nonmanipulable SDF, one must have:

Social Choices

A	A	A	A	not B	not B

Next, if the social choice in column 5 or 6 were C, person 2 would have an opportunity to manipulate in column 4. That is, he could force the choice of C instead of A, when his real preferences are $\begin{matrix} B \\ C \\ A \end{matrix}$, by pretending

his preferences are as in 5 or 6. Therefore, for any nonmanipulable SDF, one must have:

Social Choices

A	A	A	A	A	A

Similarly reasoning forces particular social choices as one drops down rows and fills out all 36 cells in Figure 3. The rest of the filling out process is left to the interested puzzle solver. When all 36 cells are filled out, the result is:

Social Choices

A	A	A	A	A	A
A	A	A	A	A	A
B	B	B	B	B	B
B	B	B	B	B	B
C	C	C	C	C	C
C	C	C	C	C	C

This is a replica of Figure 2: *Person 1 is a dictator.*

This outcome became inevitable when it was assumed that the social choice for the first row, third column cell was A. Had B been assumed, person 2 would have been the dictator.

In either case, a nondegenerate nonmanipulable SDF must be dictatorial. Therefore, there is no nondegenerate, nondictatorial cheatproof SDF. This completes the proof of the theorem.

Before finishing this section I should remark on the special case nature of the proof: Here there are two individuals, three alternatives and no indifference. Using more sophisticated tools the theorem is generalizable to two or more individuals, three or more alternatives, and indifference permissible. Original proofs are in Gibbard (1973) and Satterthwaite (1973, 1975) and two refined second generation proofs can be found in Schmeidler and Sonnenschein (1974, forthcoming).

In any case, the result generalizes: *A nondegenerate, nondictatorial cheatproof SDF does not exist.*

IV. CONCLUSION

The impossibility theorem of Gibbard and Satterthwaite is so definitive that it ought to cap a 200-year-old search for an ideal voting procedure: There is no ideal voting procedure. However, it will not stop the search. It will raise and is raising hosts of questions just as Arrow's theorem did: What non-SDF procedures (like random plurality voting) are not manipulable (Barberá (1977), Fishburn (1976), Gibbard (1977))? What restrictions might be placed on allowable preference profiles and/or ballots to escape the theorem (Blin and Satterthwaite (1976))? What about the sizes of manipulating coalitions? What happens when the number of voters is very large (Pazner and Wesley (1978)) (in which case the probability that one person's manipulation will affect the outcome is effectively zero)? What can be said about the manipulation of those more general SDF's that map preference profiles into *sets* of best alternatives, rather than singleton winners (Kelly (1977))?

The questions will go on and on, because at issue is the fundamental nature of democratic decision processes, and this issue is obviously profound. But the result that it is logically impossible to escape deficiencies in voting procedures, which first surfaced in Condorcet and de Borda, will, inevitably, remain.

APPENDIX

The notation and spirit of the proof of the proposition are borrowed from Schmeidler and Sonnenschein (1974). Let F represent an SDF, $(P_1, P_2, \ldots, P_n)$ a preference profile (for n persons). XP_iY then means person i prefers X to Y. F maps preference profiles into alternatives; we can write, for example, $F(P_1, P_2, \ldots, P_n) = X$. If this is the case for some preference profile, alternative X is said to be in the range of F.

Proposition: Suppose the SDF F is cheatproof, and X is in the range of F. If XP_iY for all i, then $F(P_1, P_2, \ldots, P_n) \neq Y$.

Proof: Define P_i' from P_i by moving $\{X, Y\}$ to the top of i's list, preserving the $\{X, Y\}$ ordering ($XP_i'Y$ for all i), and preserving the ordering among all elements other than X and Y.

First, I claim that $F(P_1', P_2', \ldots, P_n') = X$. Suppose to the contrary that $F(P_1', P_2', \ldots, P_n') \neq X$, and let $(P_1'', P_2'', \ldots, P_n'')$ be a preference profile which does give rise to the choice of X.

Define $X_0 = F(P_1'', P_2'', \ldots, P_n'') \quad (= X)$

$$X_1 = F(P_1', P_2'', \ldots, P_n'')$$

$$X_2 = F(P_1', P_2', \ldots, P_n'')$$

$$\vdots$$

$$X_n = F(P_1', P_2', \ldots, P_n') \quad (\neq X).$$

Let j be the smallest number for which $X_j \neq X$. Then

$$F(P_1', \ldots, P_{j-1}', P_j'', \ldots, P_n'') = X, \quad \text{but}$$
$$F(P_2', \ldots, P_{j-1}', P_j', \ldots, P_n'') = X_j \neq X.$$

By the construction of P_j', $XP_j'X_j$. This implies F is manipulable by j at $(P_1', \ldots, P_j', P_{j+1}'', \ldots, P_n'')$, a contradiction. Therefore, $F(P_1', \ldots, P_n') = X$, as claimed.

Next, suppose that $F(P_1, P_2, \ldots, P_n) = Y$. Define

$$\begin{aligned} Y_0 &= F(P_1', P_2', \ldots, P_n') \qquad (= X) \\ Y_1 &= F(P_1, P_2', \ldots, P_n') \\ Y_2 &= F(P_1, P_2, \ldots, P_n') \\ &\vdots \\ Y_n &= F(P_1, P_2, \ldots, P_n) \qquad (= Y). \end{aligned}$$

Let k be the largest number for which $Y_k \neq Y$. Then

$$F(P_1, \ldots, P_k, P_{k+1}', \ldots, P_n') = Y_k \neq Y, \text{ and}$$
$$F(P_1, \ldots, P_k, P_{k+1}, \ldots, P_n') = Y.$$

There are two cases to consider. (i) If $Y_k = X$, then $XP_{k+1}Y$ by assumption, and F is manipulable by $k+1$ at $(P_1, \ldots, P_{k+1}, \ldots, P_n')$, a contradiction. (ii) If $Y_k \neq X$, then $YP_{k+1}'Y_k$ by the construction of P_{k+1}', and F is manipulable by $k+1$ at $(P_1, \ldots, P_k, P_{k+1}', \ldots, P_n')$, again a contradiction. In either case the supposition that $F(P_1, \ldots, P_n) = Y$ is untenable, which completes the proof of the proposition.

REFERENCES

Arrow, K. J., *Social Choice and Individual Values*, 2nd ed., Wiley, New York, 1963.

Barberá, S., "The Manipulability of Social Choice Mechanisms that Do Not Leave Too Much to Chance," *Econometrica*, 45, October 1977, 1573-1588.

Black, D., "On the Rationale of Group Decision Making," *The Journal of Political Economy*, 56, 1948, 23-34.

_____, *The Theory of Committees and Elections*, Cambridge University Press, Cambridge, England, 1958.

Blin, J. M. and Satterthwaite, M. A., "Strategy-Proofness and Single-Peakedness," *Public Choice*, 1976, 51-58.

Brams, S. J. and Fishburn, P. C., "Approval Voting," mimeo., 1977.

De Grazia, A., "Mathematical Derivation of an Election System," *Isis*, 44, June 1953, 42-51.

Dummett, A. and Farquharson, R., "Stability in Voting," *Econometrica*, 29, 1961, 34-44.

Feldman, A. M., "A Very Unsubtle Version of Arrow's Impossibility Theorem," *Economic Inquiry*, 1974, 534-546.

Fishburn, P. C., "Acceptable Social Choice Lotteries," *Proceedings of the International Symposium on Decision Theory and Social Ethics*, Bavaria, 1976.

Gibbard, A., "Manipulation of Voting Schemes: A General Result," *Econometrica*, *41*, July 1973, 587-601.

_____, "Manipulation of Schemes that Combine Voting with Chance," *Econometrica*, *45*, April 1977, 665-681.

Kelly, J. S., "Strategy-Proofness and Social Choice Functions Without Singlevaluedness," *Econometrica*, *45*, 1977, 439-446.

Pazner, E. and Wesley, E., "Cheatproofness Properties of the Plurality Rule in Large Societies," *The Review of Economic Studies*, February 1978, 85-92.

Satterthwaite, M. A., "The Existence of a Strategy Proof Voting Procedure: A Topic in Social Choice Theory," Ph.D. Dissertation, University of Wisconsin, 1973.

_____, "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, *10*, 1975, 187-217.

Schmeidler, D. and Sonnenschein, H., "The Possibility of a Cheat Proof Social Choice Function: A Theorem of A. Gibbard and M. Satterthwaite," Discussion Paper No. 89, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston, Illinois, 1974.

_____, "Two Proofs of the Gibbard-Satterthwaite Theorem on the Possibility of a Strategy-Proof Social Choice Function," *Proceedings of a Conference on Decision Theory and Social Ethics*, Reidel Publishing Co., forthcoming.

Sen, A. K., *Collective Choice and Social Welfare*, Holden-Day, Inc., San Francisco, 1970.

_____, "Social Choice Theory: A Re-examination," *Econometrica*, *45*, No. 1, 1977, 53-90.

Vickrey, W., "Utility, Strategy, and Social Decision Rules," *Quarterly Journal of Economics*, *74*, 1960, 507-535.