Why nice guys lose: impossibility theorems in social choice theory

David A. Craven

13th October 2010

The idea of this talk is to discuss two impossibility theorems in social choice theory, proving the limits of voting systems, and in particular how they must inevitably become a (mathematical) game. We then move to discuss strategies in games in general, at a very basic level.

1 Social Welfare Functions

A social welfare function is an attempt to formalize the idea of a group of people jointly making a choice. Formally, we define it as follows.

Definition 1.1 Let $\mathcal{A}$ be a collection of alternatives, and let N be a set of voters. Let $L(\mathcal{A})$ denote the set of all total orders on $\mathcal{A}$. A *social welfare function* is a function $F : L(\mathcal{A})^N \to L(\mathcal{A})$. A point P in $L(\mathcal{A})^N$ is called a *preference profile*.

More informally, a social welfare function is a recipe for turning preferences of the members of a society into a preference for the society. Most social welfare functions are uninteresting, and we give some

Definition 1.2 Let $\mathcal{A}$ be a set of alternatives, N a set of voters, and F a social welfare function.

(i) F is a *dictatorship* if F is a projection map from one of the co-ordinates, and a *non-dictatorship* otherwise.

(ii) F satisfies *independence of irrelevant alternatives* (IIA) if, whenever $P = (P_1, \ldots, P_N)$ and $P' = (P'_1, \ldots, P'_N)$ are two preference profiles, and a and b have the same ordering in P_i as P'_i, then a and b should have the same ordering in $F(P)$ as $F(P')$.

(iii) F has *Pareto efficiency* if, for $a, b \in \mathcal{A}$, whenever $a < b$ for all elements of the preference profile P, $a < b$ in $F(P)$.

Informally, a dictatorship is one where there is a chosen individual, with preference P_i, such that $F(P) = P_i$, so that he decides the outcome. The IIA condition states that, for example, if a, b and c are the candidates, and if $a > b$ in $F(P)$, then how the voters judge c should not affect this. Pareto efficiency says that a unanimously preferred alternative should win over a less preferred one.

Later we will be interested also in tactical voting. The precise definition is as follows.

Definition 1.3 Let $\mathcal{A}$ be a set of alternatives, N a set of voters, and F a social welfare function. We say that F is *subject to tactical voting* if there exists an individual i, and preference profile $P = (P_1, \ldots, P_N)$, such that by replacing P_i with a different ordering P_i' (and writing P' for the preference profile with P_i' instead of P_i), the maximal element of $F(P')$ is greater than the maximal element of $F(P)$ with respect to P_i.

This technical definition is really saying the following: if individual i knows the voting patterns of all other individuals, then we can change his vote from his actual preference in order to make the 'winner' (maximal element of $F(P)$) more preferred to him (according to his actual preferences).

2 Arrow's Impossibility Theorem

Arrow's impossibility theorem is a result concerning the ability (or lack thereof) to produce a social welfare function with three properties. The version that we present is strictly stronger than the original version, which replaces one of the conditions (Pareto efficiency) by two others.

Theorem 2.1 (Arrow, 1950, (1963)) Let $\mathcal{A}$ be a set of alternatives with $|\mathcal{A}| \geqslant 3$, N a set of voters, and F a social welfare function. It is not possible for F to satisfy Pareto efficiency, non-dictatorship (so that $N > 1$) and independence of irrelevant alternatives.

Proof: Let a be an alternative, and choose a preference profile $P = (P_1, \ldots, P_N)$ such that a is maximal or minimal in each P_i. We claim that a is maximal or minimal in $F(P)$. Let u and l be such that $u > a$ and $a > l$ (u and l mean upper and lower respectively). Let P' be the profile obtained from P by switching u and l in P_i if $u > l$ originally. Notice that as u and l maintain their positions relative to b, by IIA $u > a > l$ in P'; however, by Pareto efficiency $l > u$, a contradiction. Hence a is either maximal or minimal in $F(P)$.

Let $P^{(0)}$ denote any preference profile in which a in a minimal alternative for each voter, and let $P^{(i)}$ denote the preference profile obtained from $P^{(0)}$ by moving a from bottom to top in the first i voters' preferences. By Pareto efficiency, a is minimal in $F(P^{(0)})$ and maximal in $F(P^{(N)})$, and so there exists $1 < i \leqslant N$ for which a is minimal in $F(P^{(i-1)})$ and maximal in $F(P^{(i)})$ (using the first argument).

We will prove that individual i is a dictator, completing the contradiction. Choose x and y to be distinct alternatives, neither equal to a. We show that $x < y$ for voter i if and only if $x < y$ for the group. Let P' be obtained from $P^{(i)}$ by placing y above a for voter i, and allowing $P^{(j)}$ to swap x and y (or not) for all $j \neq i$. Since a and y have the same relationship in P' and $P^{(i-1)}$ for every voter, and in $P^{(i-1)}$ a is minimal, $y > a$ in $F(P')$ by IIA. Similarly, since a and x have the same relationship in P' and $P^{(i)}$, and a is maximal in $F(P^{(i)})$, $x < a$ in $F(P')$. Hence $x < a < y$, and so $x < y$ in $F(P')$. As P was chosen arbitrarily, with only the positions of a fixed, in *any* preference profile P we have that $x < y$ in $F(P)$ if and only if $x < y$ in P_i.

Notice that the construction above implies that, given any alternative $z \in \mathcal{A}$, there exists an individual j such that, for all preference profiles P, $x < y$ in $F(P)$ if and only if $x < y$ in P_j, for all $x, y \neq z$. Let y be as in the previous paragraph, and choose $z \neq a, y$; let j be as in the previous sentence. Since $a < y$ in $F(P^{(i-1)})$ and $a > y$ in $F(P^{(i)})$, and only i changed his vote, we must have that in fact $j = i$; thus i is a dictator, as claimed. □

While the non-dictatorship and Pareto efficiency conditions are natural and always satisfied by reasonable voting systems, it is IIA that is the condition that is violated; the reason is generally that adding another candidate splits the vote, allowing a different candidate to win.

3 Gibbard–Satterthwaite Theorem

The natural consequence of Arrow's theorem is that, if there are more than two candidates then IIA cannot hold. The next question is whether the lack of IIA is necessarily a problem; that is is the fact that IIA doesn't hold mean that the voting system has problems. One such problem is tactical voting, as outlined earlier. The next theorem was conjectured in 1961 by Michael Dummett and Robin Farquharson. (Farquharson studied at Brasenose, then Queen's and Nuffield, before going insane.)

In this theorem, the social welfare function should simply make a choice, rather than a ranking, and so the definition of a dictator here is one where $\max F(P) = \max P_i$.

Theorem 3.1 (Gibbard–Satterthwaite, 1973) Let $\mathcal{A}$ be a set of alternatives with $|\mathcal{A}| \geqslant$

3, N a set of voters, and F a social welfare function. Write $f(-) = \max F(-)$. If F is non-dictatorial, and f is surjective, then F is subject to tactical voting.

The proof of this theorem, like the previous one, proceeds in stages. Because it is slightly more complicated, we single then out as lemmas.

Lemma 3.2 Let P be a preference profile, and suppose that $f(P) = a$. If P' is another profile such that, for all $a \neq b \in \mathcal{A}$, we have that $b < a$ in P'_i whenever $b < a$ in P_i, then $f(P') = a$.

Proof: Choose $1 \leqslant i \leqslant N$, and suppose that $P'_j = P_j$ for $j \neq i$; and let $b = f(P')$. Since F is not subject to tactical voting (NSTV), $b > a$ in P'_i (else voter i would vote as in P_i). However, again by NSTV $b < a$ in P_i, since else voter i would vote like P'_i. Thus $b < a$ in P'_i by hypothesis, and so $b = a$. The result now follows because an arbitrary P' can be obtained from P by changing one P_i at a time. □

This lemma implies that F is (sort of) Pareto efficient, in the following sense.

Corollary 3.3 If P is a preference profile and $a, b \in \mathcal{A}$ are such that $a < b$ in P_i for all $1 \leqslant i \leqslant N$, then $f(P) \neq a$.

Proof: Suppose that $f(P) = a$; since f is surjective, there exists P' such that $f(P') = b$. Let P'' be any preference profile such that P''_i relatively ranks all $x \in \mathcal{A} \setminus \{a, b\}$ like P_i does, with b maximal and a just below, for all $1 \leqslant i \leqslant N$. By Lemma 3.2, since $f(P) = a$ we have that $f(P'') = a$. Also, since P''_i has b maximal for all i, and $f(P') = b$, by Lemma 3.2 again we have that $f(P'') = b$. Hence $a = b$, a contradiction, so that $f(P)$ cannot be a, as claimed. □

Call a preference $P_i \in L(\mathcal{A})$ a *winning strategy* for voter i if $f(P) = \max P_i$ for all preference profiles P with ith co-ordinate P_i. In other words, a winning strategy is a preference such that the voter always gets his wish, regardless of other votes. Notice that a dictator is a voter for whom every preference is a winning strategy.

Lemma 3.4 If $N = 2$ then the theorem holds.

Proof: Let P_1 be some ordering of the elements of $\mathcal{A}$, without loss of generality starting $a > b > \cdots$. Let $P = (P_1, P_2)$ be such that P_2 matches P_1 but with a and b swapped. By Corollary 3.3, either $f(P) = a$ or $f(P) = b$. Suppose that $f(P) = a$. (The case $f(P) = b$ is exactly similar.) Suppose that P_1 is not a winning strategy, so that there exists P'_2 such that $f(P_1, P'_2) = c \neq a$. Since F is NSTV, we cannot have that $c = b$, else voter 2 could have

preference P_2 and vote according to P_2'. Thus $c \neq a, b$. By Lemma 3.2, we may assume that c is maximal for P_2'. Let P_2'' be obtained from P_2' by starting $b > c > a$ and keeping the rest of the alternatives in the same relative order. Write $P' = (P_1, P_2')$ and $P'' = (P_1, P_2'')$.

We claim that $f(P'') = a$. To see this, it can only be a or b by Corollary 3.3, but it cannot be b by NSTV (with P'' the vote and P the preference). However, if voter 2 had preference P_2'' and voted sincerely the winner would be a, and if he voted P_2' the winner would be c, higher on his list. This violates NSTV, and so P_1 is a winning strategy.

Hence, for every possible $P_1 \in L(\mathcal{A})$, either P_1 is a winning strategy for voter 1, or P_1 with the largest two elements swapped is a winning strategy for voter 2. If it is the same voter for all P_1 then this voter is a dictator, so both voters have winning strategies. It follows clearly that all winning strategies have the same maximal element a. Finally, if $Q_1 \in L(\mathcal{A})$ is such that $x > y > a$ is it's largest three elements, then neither Q_1 nor Q_1 with largest two elements swapped can be a winning strategy for either voter, a contradiction. Thus there is a dictator. □

We now prove the theorem. We proceed by induction on N, the number of voters. Let F be a social welfare function and let $f(-) = \max F(-)$. Let $g(-,-)$ be a voting rule given by

$$g(P_1, P_2) = f(P_1, P_2, \ldots, P_2).$$

We claim that g is surjective and NSTV, and hence either voter 1 or voter 2 is a dictator. Assume the claim.

If voter 1 is a dictator for g then he is a dictator for f, by Lemma 3.2. Hence voter 2 is the dictator for g. Let $P_1' \in L(\mathcal{A})$ be fixed, and let h be defined by

$$h(P_2, \ldots, P_N) = f(P_1', P_2, \ldots, P_N).$$

Since h has $N-1$ voters, it satisfies the theorem by induction. Since voter 2 is a dictator for g, h is surjective, and it is clearly NSTV as f is. Hence there is a dictator for h, which (WLOG) can be assumed to be voter 2 for f, and hence voter 1 for h (with preference P_2).

Finally, fix $P_3', \ldots, P_N' \in L(\mathcal{A})$ but unfix P_1, and let $k(P_1, P_2) = f(P_1, P_2, P_3', \ldots, P_N')$. As voter 1 is a dictator for g, k is onto (setting $P_1 = P_1'$ as above) Also, clearly k is NSTV as f is, so there is a dictator for k. It cannot be the first voter, since this contradicts the dictator for g, so it is the second voter. As P_3', ... P_N' were chosen arbitrarily, we see that voter 2 is a dictator for f, as needed.

It remains to prove the claim: that is, that g is onto and NSTV. By Corollary 3.3 g is surjective (let P_1 and P_2 all choose a particular alternative) so it remains to show NSTV. Suppose that g is subject to tactical voting. It cannot be with respect to voter 1, as then

f would be, so there is $P_1, P_2, P_2' \in L(\mathcal{A})$ such that $g(P_1, P_2') > g(P_1, P_2)$ with respect to P_2. However, $g(P_1, P_2') = f(P_1, P_2', \ldots, P_2')$ and $g(P_1, P_2') = f(P_1, P_2, \ldots, P_2)$; changing the voters $2, \ldots, N$ one by one from P_2 to P_2' eventually reaches a point where the voting changes from $g(P_1, P_2)$ to $g(P_1, P_2)$ (at voter i say) then voter i can vote tactically. This contradiction proves that g is NSTV, and completes the proof.

This tells you why nice guys lose: if we assume that the social welfare function treats all voters equally (and we may as well) then for every person, there is a time when being a lying scumbag helps you.

4 Game Theory and Strategies

Now that we know that any reasonable voting system is subject to tactics, game theory takes over. Voting systems are really N-person games, in which people have (in reality) imperfect information about other people's strategy.

In 1950, John Forbes Nash proved that, as long as we allow mixed strategies (i.e., people may choose between their options with certain probabilities) then a *Nash equilibrium* exists: this is a collection of mixed strategies such that no one person may improve their outcome by altering their strategy, given all others keep theirs constant.

One classic example of a game is *prisoner's dilemma*. You have two options: co-operate or backstab, and there are two players. If you both backstab you get 1 point each, if you both co-operate you get 3 points each, and if one backstabs and the other co-operates, the backstabber gets 5 points and the trusting person gets nothing. The Nash equilibrium is the backstab, because it's always the best strategy. Even if you iterate the game a fixed number N times the dominating strategy is to backstab every time.

However, if there are n players, each running d trials with every other player, it is no longer true that being a constant git is the right strategy. Indeed, suppose that there are ten people each playing the strategy "co-operate until the other person betrays me, then backstab from then on", and one constant backstabber. The backstabber gets $4 + d$ points from each of the ten players, yielding $40 + 10d$ points in total. However, each of the other players gets $d - 1$ points from the backstabber and $3d$ points from the other nine nice guys, yielding $28d - 1$ points in total. For $d \geqslant 3$ the backstabber comes out **worst**!

Computer trials with many different strategies have been run, and the best strategies seem to be along the lines of "assume people are nice, but exact retribution". This appears to be how group dynamics works in practice, or at least close to it.

We end with a brief description of a game, invented by Nash, Mel Hausner, Lloyd Shapley and Martin Shubik, called *So Long Sucker* for the general audience, or names by Nash more colourfully as *Fuck Your Buddy*.

The game is for four players, and each has seven tokens of the same (distinct) colour. There is a board allowing the formation of piles of tokens, and a cup (or something) into which 'dead' tokens are placed, from which they can never return.

Play starts with a randomly chosen player, who places a token on the board. He then nominates any other player to play next. Play continues in that at each stage a player places any token in his hand (which will in the future include other colours than his own) on the board, either on top of any pile, or starting a new pile. If he places a token on top of the same colour, that pile is *captured*, and we discuss this later. Otherwise, he may nominate any other player (including himself) whose colour is not represented in the pile just added to; if all colours are represented, play passes to the person whose colour is farthest from the top.

If the pile is captured, the person whose colour did the capture (not the player who is captured) chooses a token to be killed, and takes the rest to add to his hand. He makes the next move.

A token is a *prisoner* when held by a player other than the original owner. Any prisoner in a player's possession may be killed or transferred to any other player at any time; such transfers cannot be reversed. A player may not transfer or kill tokens of his own colour.

A player is *defeated* when he is given the move, but has no tokens (and hence is unable to play). However, defeat is not final until every player holding prisoners has refused to rescue him by transferring tokens. After defeat, the move returns to the player who gave the defeated player the move. If this should also defeat that player in turn, whoever gave that player the move will get the next turn, etc.

The defeated player's chips remain in play as prisoners, but are ignored in determining the order of play. If a pile is captured by the tokens of a defeated player, the entire pile is killed, and the move rebounds to the capturing player.

The winner is the last surviving player (after the others have been defeated), even if this player has no chips himself.

Agreements, coalitions, and so on, may be entered into or broken at any point, but all discussions must be held at the table.

The idea of the game is basically to prove that backstabbing and lying are necessary to win in certain situations. The game is set up so that eventually in order to be the winner you will have to be a scumbag.