Queen's Economics Department Working Paper No. 1384

Interpreting Arrow's Impossibility Theorem

Dan Usher
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

8-2017

Interpreting Arrow's Impossibility Theorem

Dan Usher August 15, 2017

Abstract: Arrow's Impossibility Theorem is commonly understood to invoke a dictatorship that is somehow lurking within our voting arrangements. The theorem has been described as proving that "any constitution that respects transitivity, independence of irrelevant alternatives and unanimity is a dictatorship". But the theorem is really not about dictatorship. It is more appropriately understood as being about the spoiler problem, about the possibility that the presence of a candidate who cannot win the election himself may, nevertheless, violate the "independence of irrelevant alternatives" by switching the outcome of the election between two other candidates. The theorem becomes that no electoral system is guaranteed to avoid the spoiler problem altogether, regardless of the options and regardless of voter preferences.

Key Words: Impossibility Theorem, Spoilers, Dictatorship

JEL Code: D60, D72

"Expressed in a non-mathematical way," Arrow's Impossibility Theorem is that "no voting system is fair, every voting system is flawed, or the only voting system that isn't flawed is a dictatorship."

Wikipedia, "Arrow's Impossibility Theorem"

So stated, the theorem, while not strictly-speaking wrong, is only true on the strength of a peculiar definition of dictatorship with few of the evil connotations of the term as commonly understood. A more reasonable and much less ominous interpretation is about the spoiler problem, the possibility that there is a candidate z who cannot win the election himself but whose entry into the election switches the outcome from a win by candidate x to a win by candidate y. On this interpretation, the impossibility in Arrow's impossibility theorem is not of avoiding a spoiler in any particular election – for spoilers are rare - but of avoiding the very possibility of spoilers by changing the electoral system.

As stated by Sen (2014,page 34), the theorem is

"If there are at least three distinct social states, and a finite number of individuals, then no social welfare function can satisfy ***U, I, D*** and **P**", where

U is unrestricted domain, meaning that any constellation of preferences is admissible,

P is the Pareto principle that social choice always ranks one option over another whenever everybody agrees that the former is preferable,

I is independence of irrelevant alternatives, meaning that *social choice between two options is never dependent on whether* or *not some third option is available*, and

D, commonly referred to as non-dictatorship, is that social choice is not restricted to any given person's preferences.

The theorem is proved by showing **D** to be false when **U**, **P** and **I** are true.

Sen goes on to say that "One common way of putting this result is that a social welfare function that satisfies unrestricted domain, independence and the Pareto principle has to be dictatorial. This is a repugnant conclusion – emanating from a collection of reasonable-looking axioms" (Sen, 2014, 35)

Focussing on the evils of dictatorship Sen adds that "We cannot begin to understand the intellectual challenge involved in Arrow's impossibility theorem without coming to grips with the focus on the informational inclusiveness that goes with a democratic commitment which is deeply offended by a dictatorial procedure. This is so even when the dictatorial result is entailed by axiomatic requirements that seem reasonable, taking each axiom on its own." (Sen, 2014, 31-32)

As stated by Geanakopolis (2005,page 212), the Arrow's Impossibility Theorem is that

"Any constitution that respects transitivity, independence of irrelevant alternatives and unanimity is a dictatorship"

The argument in this note is four-fold: i) the implicit meaning of "dictatorship" within the theorem is very different from and very much less threatening than dictatorship as the word is commonly understood, ii) axiom D might be interpreted as "diversity" rather than as "non-dictatorship", as meaning that the diversity of people's preferences is somehow respected in the formation of social choice, iii) the theorem should interpreted as showing, not how the axioms **U, P** and **I** are inconsistent with **D**, but how the axioms **U, P** and **D** are inconsistent with **I**, and iv) the principal take-away from the theorem is that minimizing the harm from spoilers, from candidates who affect outcomes of elections they cannot win, should be an important consideration in the design of voting arrangements. None of this is criticism of the theorem itself, which, so far as I can tell, is correct.

Consider the simplest possible example. Seven people have lunch together every day. Each day, they must choose one among three kinds of sandwiches, cheese (C), turkey (T) and ham (H). Assume, no matter why, that they cannot choose sandwiches individually, but must all have the same type of sandwich each day, though they can change types from one day to the next. If sandwiches are chosen by majority rule, first-past-the-post, voting, there are many combinations of preferences for which all four axioms, **U**, **P**, **I** and **D** - are unviolated. They are obviously unviolated when any four people's preferences are the same; by assumption, there is no dictator, leaving **D** unviolated, and there is no spoiler, leaving **I** unviolated. Most elections are like that.

Most but not all. Suppose people's orders of preference are these: Three people prefer ham to turkey to cheese, two people prefer turkey to ham to cheese and the remaining two people prefer cheese to turkey to ham. With these preferences, turkey beats ham in a head-to-head vote where cheese is not available, but cheese would be the spoiler, switching the winner from turkey to ham when all three sandwiches are available. Simple majority rule voting violates independence of irrelevant alternatives which requires that, no matter what the structure of voters' preferences, such outcomes can *never* arise.

To avoid this possibility, the seven voters might agree to a rule by which each person is entitled to choose the sandwich once every seven days, violating axiom **D** by the establishment of a rotating dictatorship in Arrow's sense of the term. One way or another, either **I** or **D** must be violated. Arrow's impossibility theorem is that this must be true of any and all ways of combining individual preference orderings into a social ordering for public choice.

A rule of social choice according to which each person gets to choose the sandwich once every seven days is a violation of axiom **D**, but such dictatorship (as implicitly defined within the theorem) is innocuous, with none of the evils we normally associate with dictatorship. "Dictatorship" invokes

images of Hitler and Stalin, images that have nothing to do with what the impossibility theorem is really about. There is no Nazi Party or Communist Party. The dictator as conceived in axiom **D** is not an evil fellow who murders people or puts them in concentration camps. He is not the *destroyer* of democratic government or even the *candidate* who wins by rigging the election. He is merely a *voter* who gets his way regardless of anybody else's preference. Typically, a real dictator holds no elections or rigs elections so that he comes out the winner; how the dictator himself votes is almost irrelevant. Putin does not cease to be a dictator if he abstains or if he chooses to vote for the opposition, as would be the case for a dictator in the world of Arrow's theorem. Nothing in the impossibility theorem indicates how the dictator is chosen, requires the dictator to be the same person from one election to the next or even guarantees that dictatorship is harmful. In the "impossibility" world, dictators get what they vote for. Actual dictators get what they want regardless of elections. There is much concern today about an apparent resurgence of dictatorship; the impossibility theorem has nothing to say about that.

A case can be made for abandoning the term "dictator" altogether, and for using the term "diversity" instead. Axiom **D** would require that social choice respect society's diversity of preference, at least to the extent that no one person's preference is destined to prevail regardless of the preferences of the rest of the community. One might think of that as the minimal requirement for *social* choice. Strictly speaking, the meaning of the axiom would be unchanged, but the repugnant connotations of the word dictator would be removed.

Axiom **I** is an extension from individual rationality to social choice. A rational person has a consistent order of preference. A person who chooses ham over turkey when nothing else is available but switches to turkey when cheese becomes available is irrational. Axiom I places a comparable requirement on public choice. Arrow describes the requirement as follows: "Suppose an election is held with a certain number of candidates in the field, each individual filing his list of preferences, and then one of the candidates dies. Surely the social choice should be made by taking each of the individual's preferences lists, blotting out completely the dead candidate's name, and considering only the orderings of the remaining names in going through the process of determining the winner. That is, the choice to be made among the surviving candidates should be independent of the preferences of individuals for candidates not in S." (Arrow, 1963, page 26). Resurrected, the dead candidate would be a spoiler if his appearance affected choice among the surviving candidates.

Axiom **I** is that there can be no spoiler, that social choice between two options is *never* dependent on whether or not some other option is available, regardless of the voting method and regardless of the constellation of preference. The word "never" is critical. A single incident is enough to violate the axiom. Axiom **I** is very strong. The axiom is that for any and every method of aggregating people's preferences into social preference – by voting or by some other means – there is no constellation of

individual preferences, however peculiar or unusual, for which the social choice between two given options is dependent on whether or not some third option is available.

Should we think of society as irrational when axiom **I** is violated in elections? Perhaps we should, but, if so, we must think of society as irrational whenever a "spoiler" affects the outcome of an election, and we must recognize that such irrationality is ubiquitous in democratic decision-making. The most cited violation is the contest among Bush, Gore and Nader in the 2001 US Presidential election where the difference in votes between Bush and Gore was much less that the number of votes for Nader and where, it is commonly believed, most of Nader's votes would have gone to Gore rather than Bush if Nader had dropped out of the race. There are more such incidences, but most elections are not like that. Dependence on irrelevant alternatives is a somewhat rare event. Independence of irrelevant alternatives means that such events can never happen, no matter what the structure of preferences or how individual preferences are combined in social decision-making.

In a recent note on how he came to discover the impossibility theorem (2014, page 144), Arrow had this to say: "The social ordering must satisfy two properties: it must reflect in some sense the preference ordering of individuals, and in making social choice from any given set of alternatives, it should use information about the preference ordering of individuals among those alternatives only. Further, it should be defined for any conceivable set of individual preference orderings, i.e., it is a functional, called the social welfare functional…..The second property above, called Independence of Irrelevant Alternatives, has the particular implication that the choice from any two-alternative set depends only on the preferences of individuals as between those alternatives." Arrow takes it as axiomatic that a process of public decision-making should inherit this feature of private decision-making, that what is rational in private decision-making should be inherent in public decision-making too. Alas, that is just not so. Though characteristic of any rational person's decision-making, independence of irrelevant alternatives is violated in public decision-making based upon individual preferences.

Following a suggestion by Ian Little, Arrow (1963, page 106) draws a distinction between "a social welfare function" and a "social decision process", where the former can be expected to preserve independence of irrelevant alternatives but the latter cannot. The social welfare function is somebody's assessment of the welfare of the nation as a whole, sometimes, though not necessarily, expressed as a "uniform income equivalent", an income which, if everybody had precisely that income, would create the same social welfare (as seen by the person whose function it is) as the actual distribution of income. The social welfare function is analogous to a person's utility function, with incomes of different people playing the role of quantities of goods, but with one mind combining

incomes of different people into a social measure. In the one as in the other, rationality of individual choice – including the independence of irrelevant alternatives - must be preserved.[1]

Social decision processes are different, in part because the social welfare function is seen differently by different people and in part because, in voting or other aspects of public decision-making, people act to promote their own interests as well as the interests, as they see them, of society as a whole. Social decision processes cobble together public decisions where there is no universally-recognized social welfare function and where the rationality to be expected from individual decision-making is more than one can reasonably expect.

Formally, the impossibility theorem is that four axioms - called ***U, I, D*** and **P –** cannot be true at once. If any three are true, the fourth must be false. In discussing the theory, it is customary to rephrase it as saying that **U**, **P** and **I** cannot all be true unless **D** is false, where D is non-dictatorship and violation of **D** means that there must be a dictator as defined within the theorem. With **D** interpreted as "diversity" (that nobody gets to determine the outcome all by himself), the theorem could equally-well be restated as showing that, together **U**, **P** and **D** require that **I** be false, meaning only that, regardless of how individual preferences are aggregated, there is always some constellation of preferences in which a spoiler would arise. This seemingly trivial restatement of the theorem has large consequences about how the theorem is interpreted in its application to actual voting arrangements. The usual formulation

[1] Imagine a person j whose perception of social welfare, W_j, is representable by average utility

$$W_i = (1/n)\sum_1^n u^j(y_i)$$

where n is total population, y_i is the income of person i and $u^j(y_i)$ is person j's perception of the utility of a person with income y_i. Person j's uniform income equivalent of the entire distribution of income is $\boldsymbol{Y}_j$ defined implicitly by the equation

$$W_j = u^j(\boldsymbol{Y}_j)$$

So defined, the uniform income equivalent is the income such that social welfare as seen by person j would be the same if everybody had that income as it is with the actual income distribution. Imagine an election in which voters' only concerns about the different candidates are for the distributions of income that their policies would provide. By assumption, person j wants people to be prosperous and happy, but he doesn't care whether the population is large or small. If all voters' utility of income functions were the same and people voted altruistically to attain the largest attainable social welfare, then the candidate supplying the largest uniform income equivalent would win the election unanimously and independence of irrelevant alternatives would be preserved. If candidate x beats candidate y when candidate z is not in the race, the entrance of candidate z could never swing the election to candidate y. Either candidate z would win, or candidate x would remain the winner. This feature of elections need not be preserved when people's utility of income functions are not the same.

suggests the death of democracy. The alternative formulation identifies a difficulty that democracies can and do live with.

The alternative formulation seems preferable for two reasons: The first might be called logical. To say that **U**, **P** and **D** require that **I** be false is to infer the possibility of a common occurrence from three generally-recognized properties of public choice. The usual formulation is convoluted. To say that **U**, **P, I** violate **D** is to combine as postulates two generally-recognized characteristics of public choice with a characteristic of individual behaviour which we know does not extend to public choice at all, and to infer from this mix a counter-intuitive inference with little relevance to voting or to any other method of public decision-making. That one postulate we know to be false implies another that is equally false is of little help in actual decision-making.

The other reason is political. The lesson in the impossibility theorem becomes not that dictatorship is inevitable (for it really has nothing to say about dictatorship), but that a major consideration in the design of voting arrangements and other methods of social choice is, as much as possible, to keep the spoiler away. Keeping the spoiler away is an important argument in favour of the alternative vote (sometimes called instant run-off elections) as compared with ordinary first-past-the-post voting and is central to Maskin and Sen's (2017) advocacy of what they call majority voting. The impossibility theorem shows that the spoiler cannot be banished completely. Spoilers may, nevertheless, be more likely to appear in some arrangements than in others.

How much does all this matter? Much depends on how the wrong candidate – wrong in the sense that most voters prefer some other candidate - is elected, whether through political machinations or by random quirks in the electoral procedure. As long as mistakes really are random, benefiting neither good guys nor bad guys on average, and are not too frequent, they may make little difference in the long run. It does not matter a great deal if a candidate preferred by 51% of the electorate is occasionally defeated by a candidate preferred by 49% of the electorate when the error is inherent in a voting system that has been in force since time immemorial and is not subject to manipulation by politicians today. The spoiler problem is closer to a random mistake. There are worse electoral diseases: gerrymandering, disenfranchisement of some class of voters to increase incumbents' chances of re-election, huge disproportions in competing parties' access to campaign funds and the not-completely-avoidable risk of a real dictator being elected. Democracy can live with a system in which the losing candidate might occasionally beat the winning candidate in a head-to-head vote, as long as this does not happen too often and as long as such outcomes are seen as essentially random.

Looked upon as showing that axioms **P**, **U** and **D** imply axiom **I** to be false, the theorem is almost obvious. What needs to be established is not that there is a spoiler in every election – for there obviously is not – but that, for every method of combining different people's orders of preference into one order of preferences for society as a whole, there is some constellation of preferences – however unlikely or peculiar – in which a spoiler may appear. It may be sufficient to note that, for every rule

amalgamating individual preferences into social preference, it is always possible that option x beats option y by a hair's-breadth when option z is not in the race, but that the appearance of option z takes away enough support for option x that the winner becomes option y instead. Axiom **I** is that this can never, ever happen. Surely, there is some constellation of preferences for which it can. Q.E.D. I leave it to the reader to decide whether this is a demonstration or a real proof.

There is also some question as to whether violation of the independence of irrelevant alternatives causes voting to be unfair as asserted in the quote from Wikipedia at the beginning of this note. If fair means nothing more than that y must never be chosen when x can beat y in a head-to-head vote, then, indeed, Arrow's impossibility theorem shows that voting may be unfair. But there is another more common meaning of fair for which the conclusion does not hold. Fair is usually interpreted to mean "in accordance with rules that have been in place for a long time" or at least since a time before anybody knew who the candidates in today's election would be, rules not chosen by the government in office today to increase its chance of winning an election today or tomorrow. On this definition of fairness, first-past-the-post voting is fair despite the fact that it may give rise to spoilers from time to time. Gerrymandering is unfair; violation of the independence of irrelevant alternatives is unfortunate and undesirable but not unfair.

****

The impossibility theorem may be interpreted as meaning that

i) The only escape from spoilers – the only guarantee that no spoiler can ever arise regardless of the constellation of people's preferences – is to assign social choice to a "dictator", to some designated person, no matter whom, or

ii) With social choice that takes account of the diversity of people's preferences, there is always some possibility – some constellation of people's preferences, however peculiar or unlikely – of a spoiled election where the winner of the election depends on the presence or absence of some other candidate who cannot himself win the election.

The first interpretation of the impossibility theorem carries the ominous, and in the end unwarranted, suggestion that democracy is deeply flawed, with dictators waiting in the wings. The second shows that no electoral system can eliminate all possibility of spoilers, that public decision-making contains a touch of what would be irrationality in private decision-making. The moral of the theorem becomes to recognize the possibility of spoilers as an unfortunate fact of life and to include minimization of the incidence of spoilers among the considerations in the design of a mechanism by which individual preferences are combined for public choice. The slide from democracy to dictatorship is an ever-present danger, but the impossibility theorem is about something else altogether.

References:

Arrow, Kenneth, *Social Choice and Individual Values*, second edition, Cowles Commission, 1963.

Arrow, Kenneth, "The Origins of the Impossibility Theorem" pages 29-42 in Maskin, Eric and Sen, Amartya, *The Arrow Impossibility Theorem*, Columbia University Press, 2014.

Geanakopolos, J., "Three Brief Proofs of Arrow's Impossibility Theorem", *Economic Theory*, volume26, 2005, 211-215.

Maskin, Eric and Sen, Amartya, "The Rules of the Game: A New Electoral System", *New York Review of Books*, January 19, 2017.

Sen, Amartya, "Arrow's Impossibility Theorem", pages 143-148 in Maskin, E and Sen, A.K , *The Arrow Impossibility Theorem*, Columbia University Press, 2014