Draft: Please do not cite or circulate.

Truth in voting[1]

Colin Guthrie King
Department of Philosophy
Providence College

> "The just is something which is according to proportion. The proportional is not only a property of monadic number, but of number in general: For proportionality is the equality of ratios." Aristotle, *Nicomachean Ethics* V.3, 1131a29–31

I.

Social choice theory since its inception in the 18th Century is a study of procedures of preference aggregation and their properties, but the epistemological background assumptions of theories of social choice have changed radically through its history. Thus with Condorcet, the theory of social choice was accompanied with background assumptions concerning the epistemic capabilities of individuals and the possibility for their collective decisions to be right and their judgments to be correct, also on moral and political matters.[2] Famously, Condorcet also argued that the majority's will can be determined, and that the will of the majority tends to be true.

But as the field of social choice theory developed, these assumptions changed. Decision theorists today are less likely to posit a will for a community, much less presume it capable of truth. Contemporary theory of voting and social choice are influenced in this by

[1] Warm thanks to David Estlund and participants in a 2016 MANCEPT section on the theory of voting for extensive comments on previous iterations of this paper. I owe further thanks to the members of my department at Providence College, who allowed me to perform several voting experiments in the course of our regular departmental procedures in 2017 and 2018, and thus helped practically inform parts of this paper.

[2] See for example this programmatic statement in Condorcet's *Notes on Voltaire* (1789), 7: "Since everyone who reasons correctly will arrive at the same moral ideas just as they will all arrive at the same geometrical ideas, it is equally true to say that these ideas are not arbitrary, but rather that they are certain and constant. They are, in fact, a natural result of the properties of rational, sentient beings; they derive from their very nature, so it is enough to suppose the existence of such beings for the propositions founded on these notions to be true; just as it is enough to suppose the existence of a circle in order to establish the truth of propositions developing its various properties".

the seminal work of Kenneth Arrow, in particular by his theorem that *no* procedure of preference aggregation among individuals can satisfy basic reasonable requirements for the proper ordering of individual preferences.[3] This has prompted reflection on the normative elements which may speak for one or another voting scheme, since the choice of a voting scheme itself is understood as a normative one, i.e. the choice of one (biased) preference mechanism among many.[4] Thus a certain part of contemporary social choice theory after Arrow has been devoted to justifying collective decision-making and democracy itself with recourse to axioms which ensure the rational structure of the set or sets of preferences (or judgments) which such procedures are taken to yield or represent.[5] Others have sought to make innovations in the institutional design of social choice procedures in order that they might better satisfy the requirements for a proper ordering of individual preferences, in particular with recourse to expanded formal and mathematical models for the design and measurement of such procedures.[6]

In this paper, I wish to make a modest proposal concerning the very criteria one employs in evaluating procedures of preference aggregation through voting. (Whether a vote, in the form of a cast ballot, expresses a real or presumed "preference" or not, will not be an object of concern: I assume that the ballot, if validly completed, is the relevant form

[3] Kenneth Arrow (1951/1963), *Social choice and individual values*, New York, 2nd edition. These conditions are: 1. Each individual voter ranks choices transitively, i.e. if voter *X* prefers A to B and B to C, X prefers A to C; 2. If each voter ranks a pair, e.g. A over B, in the same way, this is the ranking of the pair for the group of voters; 3. The group's relative ranking of two candidates is determined by each voter's ranking of this pair, e.g. if the group preference is A over B and B over C, then the relative ranking of A over C should hold, regardless of how each voter views B.

[4] As an example of both reflection on this fact and disagreement about the norms to invoke in voting, see Matthias Risse (2001), "Arrow's Theorem, Indeterminacy, and Multiplicity Reconsidered", *Ethics* 111, 706–734, and the response by Donald Saari (2003), "Capturing the "Will of the People"", *Ethics* 113, 333–349.

[5] One may refer to the work of Christian List and others in this connection; see also Hélène Landemore (2013), *Democratic Reason*, Princeton, who argues for the position that votes express judgments, not preferences.

[6] See for example (among many works by the same author) Steven Brams, *Mathematics and Democracy*, Princeton 2008. His formal analyses are motivated by an argument that approval voting, where voters may cast as many ballots as there are candidates, is a relatively superior procedure of preference aggregation, both in single winner and multiple winner contests.

in which a preference is to be aggregated in voting.[7]) My purpose is thus neither to attack or defend the legitimacy of collective decisions as truly reflecting the will of the people or tending true to the facts. I am interested rather in the specific problem of determining the normative criteria with regard to a specific question: given certain inputs in a procedure of preference aggregation through voting, which criterion or criteria should be prioritized in evaluating the relationship between these inputs and the real outputs the system yields in terms of representation?

This question is more concrete than those typically posed is a second-order normative question which arises when we realize that there several possible methods of aggregating preferences through voting, and that the choice between them must invoke some norms. At this point, given some experience with various voting schemes, we might all have our own intuitions about which scheme is most fair or advantageous for our particular interests. The question at issue here is as to the principles which should inform the selection of such norms, assuming we wish to appeal to norms beyond narrow self-interest or the interests of a particular party. As a first step in such a normative framework one could take up the central conditions for which Arrow's theorem shows that there is no completely successful social welfare function, and ask which should be prioritized.[8] But it should be noted that the relevant concept of fairness in evaluating the outcome of a vote is temporally informed, or

[7] Following Robin Farquharson (1969), *Theory of Voting*, New Haven, I further assume that a choice constitutes a vote under the following conditions, which pertain to the relation between individual choices and collective choices: 1. If a choice of one candidate does not produce the victory of that candidate, then the choice of another candidate will not produce the victory of that candidate; 2. Any individual voter's choice can be outweighed by those of the other voters together; 3. Every individual voter's choice has some weight. See Farquaharson (1969: 13–14, with note 8), who notes that these conditions break down if the number of voters is 2, and that they are consistent and independent for 4 voters and more.

[8] These conditions, as understood with respect to a voting system, are: 1. Non-dictatorship (preferences cannot be set by a single voter); 2. Universality (preferences must be expressed by a unique and complete ranking of choices); 3. Independence of Irrelevant Alternatives (aggregated preferences between x and y must mirror individual preferences between x and y); 4. Monotonicity (a ranking in favor of a candidate should not result in that candidate being ranked lower in the aggregation of all preference profiles); 5. Non-imposition (every possible aggregation of all preference profiles should be mappable to some set of individual preference profiles).

retrospective.[9] And in such a context of retrospective judgment, "mere conformity of an outcome to a rule aggregating votes still does not capture the idea of retrospective fairness", as that idea has as its object the right causal relation between the votes and the outcome.[10] This holds true also when we think of the outcome in terms of actual representation. The relevant sense of "truth in voting" which is at issue here, then, is coming to a better understanding of this causal relation, and how certain outcomes can more or less reflect the votes that preceded them.

In taking this approach I do not address the claim that the inputs or outputs in procedures of preference aggregation through voting represent any sort of truth *independent* of the procedures themselves. I am not arguing, for example, that more inputs will reflect more about the world, as some have argued with a view to Condorcet's Jury Theorem.[11] I hope rather to defend a claim concerning the primary political function of voting – namely: representing voters – and the norm we should apply in evaluating inputs of a voting systems and outputs (in terms of representatives). Such a norm is particularly important for evaluating electoral systems, but ultimately of central importance for the evaluation of preference aggregation procedures generally. The norm may be expressed as the following criterion: *A procedure and in particular a voting system is more faithful, i.e. has higher fidelity, to the degree in which its outputs correspond to its inputs, or: procedures with outcomes which are more faithful to the inputs are more "true" to those inputs.*

[9] This point is argued forcefully by David Estlund (2008), *Democratic Authority*, Princeton, in Chapter IV: "The Limites of Fair Procedure", in particular 69–76.

[10] Estlund (2008), 75.

[11] On this general tendency, see again Estlund (2008), Chapter XII: "The Irrelevance of the Jury Theorem", 223–236.

II.

As advertised above, we wish to explore what it means for an output to "correspond" to an input, and for an electoral outcome to be "faithful" to the votes cast, in terms of a causal relation between votes and representatives. As a first approach to this relation we may begin with a hypothetical example of a voting system for a single member district in which the majority or a plurality is determinative of the outcome, but in a manner different from that which is accustomed. In this system, the candidate with the most votes always loses and the candidate with the least votes always wins. We further suppose that the voters do not know this; otherwise, they could in fact successfully express a preference by voting for the candidate they prefer the least. Let us therefore assume for the sake of the example that the vote-counting machinery in this electoral system has been hacked, and the hackers have systematically set up the machines to declare the lowest vote-getter the winner.

Some might object to the description of such an arrangement as an electoral system, i.e. a system of voting. But clearly this arrangement would provide a mechanism for tallying votes. It satisfies three intuitive requirements for a vote formulated by Farquharson (1969).[12] It is plausible to assume that the choice of any one candidate will not lead to the election of that candidate, and that this holds for all others. Any individual vote can be outweighed by the votes of other voters together. And every individual vote has some weight. So even if the actual, hacked outcome of the vote is derived from a manipulation of the real outcome, the real outcome is in a causal and determinative relation to the actual, hacked outcome. But it is not the right causal relation: there is an asymmetry between the cause and the effect, and the information contained in the original vote has been subject to an interpretation which runs exactly counter to the intention of those who participated in the vote. For on our

[12] See note 7, above. The conditions only hold jointly for groups of voters of four or more, and do not hold for an electorate of two individuals.

scenario, the voters expect that the winner of the procedure is the person who really won a majority or plurality of the vote.

Would such a voting arrangement be able to qualify as a system of representation? Presumably, some voters will have a candidate as representative for whom they voted (assuming none of the candidates receive zero votes and there are no ties for last place in the original tally). Their votes will be part of the causal story of the election of their candidate, just in a different way than they presumed. This is important. Let us call the scenario we have described the *hacked system*, and compare it with another, similar arrangement, to be called the *disapproval system*. In the disapproval system, we make the inversion of the real outcome publicly known. This would lead the participants to change their voting behavior in such a way as to direct them to vote for those candidates of which they least approve. This kind form of disapproval voting might seem strange, but its properties are just like the first-past-the-post system with which most people in the UK and US are familiar. Unlike the hacked system, the disapproval system would be causally transparent to the participants. The hacked system in which, unbeknownst to the voters, the lowest vote-getter wins has the distinguishing feature that it is causally opaque to those who participate in it. The hacked outcome is a false representation of the aggregated votes of the voters. Its outcome is false because it is in the wrong causal relationship to aggregated votes. And the false representation of such an outcome undermines, in some way we must describe more precisely, the legitimate authority of a representative selected upon its basis.

What exactly does the hacked system fail to represent? It would seem to fail to represent the relationship between the minority and majority of votes (if one there be), a failure occurring by in fact inverting them. The disapproval system of voting in which those who get least votes win (and voters know that) is more faithful to the inputs in the following respects:

1. It is derived from a procedure which is known and (we assume) understood. Let us call this criterion the ‘epistemically transparent’ criterion.

2. The procedure makes it possible (though does not guarantee) that the *expression* of a preference in a vote will lead to an output with the right causal relationship to that preference. Let this count as the ‘causally transparent’ criterion.

3. The majorities and pluralities generated by the inputs are in the right causal relationship to respective majorities and pluralities in the outputs: there is a salient symmetry between votes and representatives. Let this be the ‘numerical symmetry’ criterion.

Let these criteria be assumed as the minimal fidelity conditions for the aggregation of preferences through voting (even if, as we have mentioned above, no system of preference aggregation may be considered perfectly faithful). We shall attempt to test and refine them; the first step will be to confront them with voting as practiced.

What, on these criteria, should qualify as a low fidelity voting system? One plausible candidate is a gerrymandered system of voting in which, due to the niceties of districting, a supermajority of votes is required to generate a modest majority in the electoral outcome. Another is a system in which a small advantage in votes generates a supermajority in electoral outcomes. Real examples of both forms of low fidelity abound, but let us take just one recent example of the latter. Let us consider the 2015 UK general parliamentary election.[13] In this election, 30.4% of a national voting population vote for Party A, 36.8% vote for party B, 12.6% vote for Party C, 7.9% vote for Party D, 4.9% vote for Party E, 3.8 % vote for party F, and the resulting distribution of seats in a legislative body is as follows:

Party A = 35% of seats in Parliament (with 30.4% of the national vote)
Party B = 51% of seats in Parliament (with 36.8% of the national vote)
Party C = 0.002% of seats in Parliament (with 12.6% of the national vote)
Party D = 0.01% of seats in Parliament (with 7.9% of the national vote)

[13] I have taken this data for this election from *The Economist*, “Square pegs, round hole”, May 8, 2015. The 2.5% of the vote for “other parties”, represented in parliament by 19 seats, were left out for the sake of simplicity. The percentage of votes is measured in absolute number of ballots cast for a member of a party, the percentage of seats is measured by the number of directly elected representatives in the House of Commons of each party.

Party E = 8 % of seats in Parliament (with 4.9% of the national vote)
Party F = 0.002% of seats (with 3.8% of the national vote)

One can see that the proportion between percentage of the national vote and seats in national legislature is not uniform for each of these six parties. Party A is Labour, Party B is the Conservatives, Party C is UKIP, Party D is the Liberal Democrats, Party E is the Scottish National Party, Party F is the Green Party. Clearly the existing arrangement is beneficial for the SNP, the Tories, and Labour, and very bad for every other party. Saliently, the system of voting (i.e. of inputs) and the rules of tabulating outputs for these inputs (i.e. the tabulation of seats in Parliament based upon outcomes in districts) have effectively translated a plurality of inputs (votes) for Party B into a majority of outputs (seats in Parliament) for Party B. This characteristic may actually be considered a virtue of the system for purposes of governability (it is a reason often forwarded in defense of such arrangements); but, at least taken as a representation of the total national vote, it arguably violates the condition that there be a salient symmetry between votes and representatives (condition 3, above) – though of course the system may, by district, fulfill all of the three minimum fidelity criteria.

It seems intuitive that voting arrangements which perform less well in fidelity of outputs to inputs are less preferable – at least in this one respect. But as is well known, not all institutional designers of modern democracies accepted fidelity as an indefeasible value. Another well-known low-fidelity voting mechanism is the institution of the Electoral College in the United States, by which the votes cast for president by citizens determine not the president herself but a number of electors, who in turn elect the president.[14] These electors are determined in advance by the respective parties and, if elected to represent the vote of their constituencies, they are bound to honor the preference expressed by the simple majority of votes cast. But they can vote in a way which does not correspond to the majority of the

[14] For a short history this institution see Robert A. Dahl (2003), *How Democratic is the American Constitution?*, New Haven, 2. Edition, 73–89.

preferences expressed in their respective state; and such electors, who cast a vote for a candidate other than the one for whom they pledged to elect, or who cast no vote at all, are called "faithless electors".[15] This language suggests the particular demand that a system of election be faithful in its outputs to its inputs: any arrangement or circumstance in an electoral system which fails or even prevents the effective transmission of preference inputs to electoral outputs is a source of distortion in precisely the sense that it makes the system less faithful or, as we shall argue, less true to its inputs. Notwithstanding this rather unusual case: The US Electoral College system for presidential elections may also be considered "low fidelity" in another sense. It almost invariably serves as one of those arrangements which purposefully amplifies outcomes from certain constituencies while muting those from others, as, infamously, in the presidential elections of 2000 and 2016, when losers of the popular vote won the electoral college.

Faithfulness of outputs to inputs may be considered a defeasible political value even in electoral systems which place much emphasis on the authority conferred by a majority and relative egalitarianism of representation across geographical regions. Some of these systems (like the British one, cited above) are so designed as to greatly privilege the party which secures a majority, going sometimes so far as to provide mechanisms which fracture the opposition, i.e. those parties which fail to gain a majority.[16] Such disproportional representation is typically justified in terms of effectiveness: it is assumed that a weak or merely tolerated "relative majority" (i.e. <50%: a plurality, not a majority) would be systematically hindered in making law and forming policy. This concern is cited in the creation of amplification systems which turn pluralities in the electoral inputs into real (i.e. >50%) majorities in

[15] Thus a "faithless elector" is only an elector who has previously made a pledge to vote for a certain candidate; unpledged electors cannot become faithless.

[16] See the discussion of these systems in Michael Dummett, *Principles of Electoral Reform*, Oxford 1997, 20–28.

electoral outcomes. This tendency can be characterized (with Dummet (1997), 22) as the defining feature of "Winner Take Most" (WTM) systems.

But systems of proportional representation (PR) also feature constraints on fidelity of inputs into outputs. Most common is the threshold above which parties must come in order to be represented in parliament, as for example in the German dual-vote system for Bundestag, where parties must exceed a 5% party vote threshold to be represented in parliament. An often cited reason for this arrangement is that it prevents "fringe" parties from entering parliament; yet it is clear that the notion of "fringe" is merely quantitative. A further, and perhaps more operative, purpose of this mechanism is to keep the number of elected parties to a manageable minimum. The 5% hurdle may be seen as a threshold for a party to be representative, in some way, of the electorate; but governability reasons also play a role in the enforcement of this threshold.

Given these institutional constraints on the fidelity of electoral inputs to outputs in the two main types of electoral systems, WTM and PR, a series of normative questions arises. Under what conditions should fidelity be a determining norm in the evaluation of electoral systems? Should further considerations such as stability, efficiency, and the maintaining of basic human rights override the norm of fidelity? And what does fidelity require exactly?

III.

In our first approach to defining fidelity we identified three minimum fidelity criteria for the output-input relation of a voting system. We may evaluate a voting system in terms of its epistemic transparency, its causal transparency, and the numerical symmetry between the sum of inputs (votes) and outputs (seats in parliament). We will later consider how fidelity in the sense of these three criteria can be an over-riding value in the evaluation of electoral systems. This is an important normative question, as both the political theorist and political practitioner may reasonably rather be concerned with the best form and functioning of government. From this point of view, one might tend to evaluate electoral procedures with a view to those things which will be conducive to good and stable government. After all, a primary concern of those actually involved in governance is that the electoral system used, whatever it be, yield outcomes which will (by and large) result in good governance. Justifying the normative value of fidelity as both a set of criteria for electoral systems and a political norm is the end-goal of our argument. First, however, we must further specify what it would mean for a voting outcome to be true or faithful to the inputs which caused it. That is what I shall try to do in this section.

There is a famous and recalcitrant problem in the theory of voting which would seem to show that the idea of fidelity in a voting system is ultimately incoherent. The problem consists in finding criteria to select the winner in an election in which majority preferences are intransitive. It can occur that a majority of voters prefers candidate A to candidate B, another majority of voters from the same voting population prefers candidate B to candidate C, and yet another majority prefers candidate C to candidate A. In this case, the transitivity of preferences which is reasonably supposed in the case of individuals is clearly violated. Under these circumstances, what could possibly count as being "true to the inputs"?

This question has elicited an abundance of theoretical responses as well as practical attempts to devise systems which will ensure that a most representative plurality obtains and determines the outcome of the procedure. Many authors would maintain that the minimum requirement of an electoral system is that it ensure that the Condorcet winner – the candidate who would beat any other candidate in one-to-one comparison – actually wins. (As virtually all voting theorists concur, many systems in use, in particular the first-past-the-post system of the UK and US, do not reliably fulfill this criterion.) But when there fails to be a Condorcet winner, who should win? One well-known approach to this question looks to the relative position of candidates in order to determine which candidate is most preferred. If there is no candidate who beats every other candidate in direct comparison, one can still see which candidate is *most* winning in comparison to other candidates (even if she is not winning in comparison to all). This is the Copeland method for determining an electoral outcome. In it, each candidate is assigned a score based upon the number of candidates above whom she is ranked by a majority. The candidate with the highest score wins.

Copeland supplements Condorcet. The main rival to the Condocret criterion is the Borda count, which will, under certain circumstances, select as winner a candidate who is not the Condorcet winner. Borda scores candidates off the ballot by assigning values to the position a candidate has on a ballot consisting of ranked choices. This method tends to select those candidates which are least controversial, whereas the Condorcet criterion will often select those candidates who are also successful under the plurality system – including those who are highly polarizing. To illustrate this using a simple example[17] from a four-way electoral race, let us assume the following distribution of ballots (with preferences expressed ordinally from left to right, where left is most preferred, right is least preferred):

Set 1: 19,000: {ABCD}
Set 2: 14,000: {BCDA}
Set 3: 2,000: {CABD}
Set 4: 1,000: {DCBA}

By the Borda count, we assign values according to the number of candidates before whom each candidate is ranked. Set 1, for example, yields for A a score of 3 x 19,000 = 57,000, Set 2 yields for A zero, Set 3 yields for A 4,000, and set 4 yields zero points for A, resulting in a total score of 61,000. Following the same procedure for the other candidates, we arrive at the result:

A = 61,000
B = 83,000
C = 55,000

[17] This example is adapted from Dummett (1997), 73.

D = 17,000

Thus, according to Borda count, we have the social preference profile {BACD}, and **B wins**. If we take the Condorcet method of simply comparing the candidates in head-to-head competitions, we arrive at a different result:

Number of voters who rank A over B (across all sets) = 21,000
Number of voters who rank B over A (across all sets) = 15,000
Number of voters who rank A over C (across all sets) = 19,000
Number of voters who rank C over A (across all sets) = 17,000
Number of voters who rank A over D (across all sets) = 21,000
Number of voters who rank D over A (across all sets) = 15,000

This yields the social preference profile {ABCD}. Since A beats every other candidate in head-to-head competition, **A wins.**[18]

Anyone who claims that there is truth in voting must confront this discrepancy (we will leave the cycle of majority preferences out of consideration for the moment). It would seem, to adapt a phrase from early analytic philosophy, that the meaning of an electoral outcome consists in the method of its verification (the preference aggregation rule employed), which just pushes the problem back to an interpretation of what a voting input or act of voting is supposed to express. The example above features one, largest group (Set 1) with a strong preference for a certain candidate (A), whereas the smaller groups all tend to put A at the bottom of their ranking (with the exception of Set 3). This may be construed as reflecting something about A: A is a tendentially polarizing candidate. Candidate B is runner-up to A in the simple plurality system, but is placed last in none of the sets. This can

[18] In this example, A also wins using the single-transferrable vote or Hare system, which progressively eliminates the lowest vote-getters in the first position of a ranked ballot of more than two candidates. A also wins using the Copeland method, with a Copeland score of 3 (A wins over 3 candidates) to B's Copeland score of 2 (B beats both D and C in head-to-head comparison) and C's score of 1 (C beats D in comparison across all sets).

be construed as reflecting that, for these four fictional groups of voters, B is a generally acceptable candidate.

This difference is one which the simple plurality or first-past-the-post system fails to represent: information concerning the acceptability of the respective candidates *in negative terms*. Those who view the majority or plurality in positive consensus as the legitimizing factor for representation will claim that the truth of voting behavior, what the vote *means*, is elicited by the simple plurality system and (in a different way) by the Condorcet and Copeland methods. Those who believe that the over-all acceptability and approval of voters, also in negative terms, is the legitimizing factor, will prefer the Borda method, or may also appeal to any number of approval-based systems of voting which are in use.[19]

The conflict of criteria is often characterized as a conflict between majority and over-all preference in the interpretation of voting inputs. In fact, both the Condorcet criterion and particularly the Borda count may elicit a kind of majority neglected in the simple plurality system, namely the majority of those who *do not* prefer a certain candidate, which we may call a negative majority. The simple plurality system is famously prone to selecting candidates as winners who are liked least by the greatest number, but who nevertheless obtain a plurality – either due to the fact that the opposition to them is split, or because they are in the middle of two candidates, each of which are supported only by a small and "extreme" group. In these cases, the simple plurality system fails to correspond to the preferences of the (negative) majority.

If one places value on fidelity to this kind of majority, then the Borda count seems to be superior even to the Condorcet criterion, as this method is more responsive to negative

[19] In approval voting for candidates in multi-candidate elections, voters have as many votes as there are candidates, who are not ranked. This system can however also be combined with ranked-choice voting. Brams (2007), Chapters 1–3, discusses various forms of approval voting for electing a single winner. The strategy here is to enrich the inputs in order to come up with a more precise method of aggregating and ordering preferences.

preferences. Borda also allows for selection of a clear winner even in cases in which the preferences of majorities cycle, as in the following example:

Set 1: 13,000: {ABCD}
Set 2: 11,000: {BCDA}
Set 3: 9,000: {CABD}
Set 4: 10,000: {DCBA}

Using the Condorcet criterion as expanded by the Copeland method, we get the following result:

Number of voters who rank A over B (across all sets) = 22,000
Number of voters who rank B over A (across all sets) = 21,000

Number of voters who rank A over C (across all sets) = 13,000
Number of voters who rank C over A (across all sets) = 40,000

Number of voters who rank A over D (across all sets) = 22,000
Number of voters who rank D over A (across all sets) = 21,000

A has a Copeland score of 2 (A beats B and D but not C).

Number of voters who rank B over C (across all sets) = 24,000
Number of voters who rank C over B (across all sets) = 19,000

Number of voters who rank B over D (across all sets) = 33,000
Number of voters who rank D over B (across all sets) = 10,000

B has a Copeland score of 2 (B beats C and D but not A).

Number of voters who rank C over D (across all sets) = 33,000
Number of voters who rank D over C (across all sets) = 21,000

C has a Copeland score of 2 (C beats D and A but not B).

Thus we have a cycle of majority preferences. According to the Borda count, the result of this election would be as follows:

A = 57,000
B = 52,000
C = 72,000
D = 41,000

This yields {CABD}, **C wins** handily. This is because C is relatively close to the top in each preference set but one (set 1), whereas the second-place candidate A is among the top two candidates only in two sets (sets 1 and 3) and last in the others (sets 2 and 4). Note that, if this were a first-past-the-post election where voters' first choice determined the outcome, C would be in *last* place.

What may one conclude from such theoretical examples? We may derive from them at least two ideal types of fidelity which can and often do conflict. One type of voting system – first-part-the-post or simple plurality – typically searches for and represents one kind of consensus, the consensus of *positive approval*. Another type of voting system for a single winner, rank-choice-voting, will tend to select the candidate who is *disliked least*, particularly when used with the Borda count rule of preference aggregation, which will prevent any candidate from winning against whom a majority of voters is opposed. As our elementary examples have shown, the Borda count fulfills one of our fidelity criteria slightly better than the Condorcet criterion: Borda better preserves negative plurality, or a plurality of negative preferences, by preventing any candidate for whom such a plurality obtains from winning. The Condorcet criterion can fail to translate this negative plurality in the inputs of a positional voting procedure into an outcome, by selecting (under certain circumstances) a winner who is nevertheless more dispreferred than another candidate.[20] Conversely, the Borda count could be said to fail to translate the positive approval pluralities which are generated by the Condorcet criterion and its extension through the Copeland method.

[20] Here I am assuming that the Borda count in its standard application adequately expresses the relative intensity of the preferences of the voter. One could contest this: interpretations of the Borda count have been introduced which weight positions, giving (say) twice as many points for being in first place as for being in second place on a ballot. Such an adaptation of the Borda count tends to approach a system in which the Condorcet winner always wins.

An argument in favor of one of these rules of preference aggregation is beyond the scope of this paper. As has been shown by Matthias Risse, it is fallacious to judge one of these aggregation methods against the standards inscribed in the other: one would have to determine standards to which any preference aggregation rule must be held.[21] What can be shown with these relatively simple examples is just that both the Condorcet method and the Borda count respect negative pluralities and a negative majority in elections with three candidates or more, which makes them both superior (by our criteria) to a simple plurality system in terms of fidelity. But the difference between these preference aggregation rules and the ideal types of fidelity we have identified in them force us to consider what else, if anything, will justify a preference for a voting system on the grounds of fidelity.

IV.

To ask which of these ideal types of fidelity is the most important one will ultimately involve an appeal to something which will either be essential or central to fidelity, or which will be something other than fidelity. If we appeal to another norm in explicating fidelity and making the claim that this or that electoral system or preference aggregation rule is more faithful, then faithfulness and fidelity (I take them to be synonymous) are no longer the legitimizing value in question. The norm which seems most relevant here – either as a further justifying norm for the notion of fidelity, or a core aspect of the notion – is fairness, in particular procedural fairness in representation. But before we engage in a normative discussion of the voting systems treated above, it is important to make the scope of our previous argument clear and distinguish a further possible sense of fidelity with respect to voting systems.

The analysis here is limited to electoral systems as a means of aggregating votes and electing a candidate in a single-winner situation with more than two candidates based upon

[21] See Matthias Risse (2005), “Why the count de Borda cannot beat the Marquis de Condorcet”, in: *Social Choice and Welfare*, No. 25/1, 95–113.

those means. For this reason it has been natural to treat votes as inputs and electoral results as outputs, since this relationship has been the focus of discussion. I have neglected to discuss the possibility of tactical and insincere voting, as these are behaviors which may influence inputs based on an expectation of outputs, and thus are extraneous to a consideration of the symmetry or asymmetry between inputs and outputs *per se*. The likelihood and influence of such behavior are indeed important factors in evaluating (i.e. normatively judging) a voting system and its fidelity to the voters' *real* preferences. But in the discussion so far, I have attempted to be agnostic about these real preferences and have focused upon fidelity as a relation between inputs and outputs, tacitly assuming that the inputs will be expressions of the voters' real preferences. This assumption will sometimes be false, but it is also hard to falsify; and in any case, it has not been essential to the argument thus far, which has concerned fidelity as a relationship between inputs and outputs (and not between real preferences and outcomes).

When we consider this further, strategic dimension of voting, the importance of fidelity of outputs to inputs becomes clear. The study of the strategic aspects of voting has brought the theory of voting within the scope of game theory.[22] As every procedure of preference aggregation is imperfect, every procedure is subject to manipulation: and this has been shown precisely for the kind of procedures under consideration, for ranked choice voting procedures with a single winner and at least three candidates.[23] A related concept which was developed before this had been proven is the notion of "straightforwardness". A voting procedure is straightforward when a voter can pursue her course of strategy with absolute

[22] For a seminal application of game theory to the theory of voting, see Farquharson (1969).

[23] This has been proven by the Gibbard-Satterthwaite theorem, which shows that every single-winner ranked choice voting method can be manipulated. This means that, for all such systems, situations obtain in which a voter can obtain a more preferred outcome by voting in deviation from her real preferences. For further discussion of this theorem (and an elegant proof for it), see Michael Dummett (2005), "The work and life of Robin Farquharson", in: *Social Choice and Welfare*, Vol. 25, No. 2/3, 475–483.

confidence that it will give at least as desirable an outcome as any other strategy.[24] The voter may fail to obtain what she wants, but if so, it will not be due to an error in choice of strategy. A strategy is straightforward if it optimizes the voter's preferences; and a procedure which offers voters only one straightforward strategy is itself straightforward. I will unpack this notion in the following.

The notion of straightforwardness helps us appreciate why the first-past-the-post or simple plurality system retains a certain attractiveness, even when it is shown to be a low-fidelity system when more than two candidates are involved. The system seems to be straightforward, and by seeming straightforward it also seems to be faithful, i.e. to have fidelity to the voters' real preferences. In fact, in elections for a single winner with only two candidates (as is often, but by no means always, the case in e.g. US Congressional districts), the procedure is straightforward in the technical sense: there is no situation in which a voter could maximize her preferences by voting in a way which differs from those preferences (whatever they may be).[25] This voting system seems to be faithful, since the relationship between inputs and outputs is epistemically transparent.

However, as is well-known from the context of US Congressional elections, further factors in the setting of a voting situation can distort the relationship between the inputs (in terms of state-wide and national voting percentage) and the outputs (in terms of legislative seats). Gerrymandering contributes to the lack of fidelity in voting situations which are straightforward in the sense determined here. For even if the voter would not have any incentive to vote in a way other than her true preferences dictate, the minority voter in a

[24] Farquharson (1969), 30.

[25] Voting for two candidates by simple majority is just a case of the binary decision rule, and its non-manipulability follows from the property of monotonicity which this rule features. The property of monotonicity can be illustrated by example: if a voter prefers A to B and B were to win without the voter's input but A would win with the voter's input, then the voter's input of A (or B) must always be determinative of the outcome. A monotonic system is one in which it will not harm a candidate's chances of success by up-ranking the candidate, nor help a candidate's chances of success by down-ranking the candidate.

gerrymandered district will experience some degree of numerical asymmetry between the percentage of vote in her district and that of her state. I submit that this makes the relationship between her vote and that of the output in terms of representation remains causally opaque because of the asymmetry between the causal power of her vote is unequal to the causal power of the votes of others in her state or country. This asymmetry and its causal effect has been theorized in terms of an "efficiency gap".[26] The "efficiency gap" is a measure of partisan symmetry which represents the number of votes "wasted" in proportion to total votes cast. On this approach, a vote is "wasted" in one of two ways: either by being cast in support of a candidate who does not win ("cracked" votes), or by being cast for a candidate in excess of the margin required to win ("packed" votes).[27] The metric provides a means for expressing symmetry and asymmetry in the proportions of votes to seats in a given district. But what is presumably at issue is not merely this numerical asymmetry, but the causal "efficiency gap" which attends it. The metric gives this causal distortion feature numerical expression. As the authors of this metric themselves state, the metric is designed to measure a causal relation targeted by partisan gerrymander: "to win as many *seats* as possible given a certain number of *votes*".[28] The normative aspect which the authors attach to this notion appeals directly to a certain causal inequity of votes in gerrymandered districts in relation to seats: "After voters have decided which party they support, … the votes cast by supporters of the gerrymandering party translate more effectively into representation and policy than do those cast by the opposing party's supporters".[29] This is identifying a truth-as-fidelity feature of a voting system and its normative implications, which were recognized in a general form already by Aristotle when he identified as a property of justice being

[26] Nicholas O. Stephanopoulos and Eric M. McGhee (2017), "Partisan Gerrymandering and the Efficiency Gap", *The University of Chicago Law Review* 82, 831–900.

[27] Stephanopoulos and McGhee (2017), 849ff.

[28] Stephanopoulos and McGhee (2017), 850.

[29] Stephanopoulos and McGhee (2017), 852–853.

"according to proportion", and this being a property not just of numbers but of the relations they express (*Nicomachean Ethics* V.3, 1131a29–31). The efficiency gap provides one plausible metric for the precise expression of the normative proportion to be applied to elections for legislative bodies.

As we have seen, even in a non-gerrymandered district, when the voting situation includes more than two candidates, first-past-the-post is neither straightforward nor does it reliably capture central features of fidelity. The symptomatic distortion features of this system with more than three candidates – split opposition and the weak middle candidate, both of whom will be dispreferred by a majority – make this system faithless in two respects. It is causally and epistemically opaque incapable of reflecting negative majorities. What is more, this system is most subject to other forms of manipulation, e.g. through strategic districting (i.e. gerrymandering) in order to maximize the electoral outcomes for a certain party or to minimize the electoral influence of certain groups. In the US at least, this electoral system has had the more recent historical effect of concentrating power in two major parties, further limiting the possibilities of representation. More recent critiques of a "tyranny of the majority" in this context have been misconstrued as a challenge to the legitimacy of binary decision procedures; in fact, this critique contains a correct analysis of the shortcomings of a system in which one is constrained to binary decision procedures in elections.[30]

Be that as it may: though empirical considerations of the effects of electoral systems are important and indeed necessary for their evaluation, we are concerned here with the problem of precisely identifying the property of fidelity in voting systems and estimating its normative value. Now we have identified two ways for truth-as-fidelity to obtain for a voting system:

> **1. Representational fidelity**, which obtains through maximum adherence to the three minimal fidelity criteria articulated above: 1. The voting scheme is

[30] See Lani Guinier (1994), *The Tyranny of the Majority: Fundamental Fairness in Representative Democracy*, with a foreward by Stephen L. Carter, New York.

> derived from a procedure which is known and (we assume) understood: it is epistemically transparent. 2. The procedure is causally equitable and transparent: voters have relatively equal power to influence the outcome through their vote. 3. There is a basic proportional symmetry between votes and seats.
>
> **2. Preference fidelity**, which ensures that a voting procedure is faithful in the sense that there is no situation in which strategic voting would have assured the voter a more preferential outcome.

Note that preference fidelity concerns votes and preferences, whereas representational fidelity concerns a relationship between votes and seats, not votes and preferences. This is important. As shown independently by Gibbard and Satterthwaite, preference fidelity is impossible to attain in a non-dictatorial voting system with more than two candidates. Preference fidelity is, in many of the voting situations we have considered, ideal and never real. Representational fidelity, by contrast, expresses scalable conditions which can be fulfilled in different ways and even with different outcomes, as we saw with regard to the possible conflicts of the Borda count and Condorcet method. On this type of fidelity, as we have seen, an electoral outcome is faithful to its inputs if it fulfills three criteria: the rules by which the outcome was derived from the inputs are epistemically transparent to those who participated in voting; that there is basic causal equity among the votes of the various voters; and there is basic numerical symmetry between votes and seats. One could have representational fidelity without preference fidelity, and we can have preference fidelity without representational fidelity. For example, in a given US Presidential election without a third-party candidate, one might appeal to preference fidelity in justifying the propriety of a given result, though representational fidelity is not given even when the Electoral College makes this voting procedure low-fidelity even when a majority of the national popular vote is garnered by the winner of the election. And saliently, when a third-party candidate is involved, neither preference fidelity nor representational fidelity are given, as was illustrated in the US presidential elections of 2000 and 2016.

V.

In conclusion, I should justify why fidelity in the sense defined here should be an overriding norm in the evaluation of voting systems. There purport to be systems which are both fair and faithless, as e.g. the Electoral College system, which was deemed appropriate by the Framers as a guarantor for the equitable representation of less populated areas in the selection of competent representatives for the Executive Branch of government, and was only later subverted to represent, albeit very imperfectly, the popular vote. Why fidelity, when the purpose of democratic government is not merely some procedural good, but to secure the basic goods of good government? Even granted that there is truth in voting of the sort we have tried to argue for, why should it be a priority?

We have seen an example of a voting system which is completely faithless, the "hacked" voting system referred to above in which the real result is always inverted and the recipient of the lowest number of votes always wins. The fidelity criteria which emerged from our analysis of this example – causal transparency, epistemic transparency, and numerical symmetry – suggest that the fairness of a voting procedure will consist in providing settings which ensure that voters have roughly equal causal powers in a given voting situation, regardless of their partisan preferences, causal powers about which they understand. This has implications for the general acceptability of such procedures. The acceptability will flow not so much through the winning of such a voting procedure as through the voting procedure having these properties. A system which undermines the causal power of voters through gerrymandering or other means also undermines an important source of its acceptability (even if it retains normative authority, as it is endorsed by certain existing institutions).

The conceptualization of voting as the exercise of a causal power has further implications for the institutional design of voting systems. If voting is an attempt to causally influence an outcome, then it is equitable to design the system in a way which diversifies the array of possible outcomes. As we have seen, not all systems perform equally well in this respect. At least in some institutional settings, first-past-the-post severely limits the diversity of outcomes and limits the expression of negative preferences, which are just as relevant for the choice of a representative as positive ones. The truth of voting which is captured in the criteria of representational fidelity will be better expressed by ranked-choice voting procedures.

From the broader perspective of a normative theory of democracy, systems which are causally and epistemically opaque, and which feature significant input-output asymmetry, may be deemed less faithful in the sense that they conceal real information about the approval of candidates. There is an intuitive connection between the faithfulness of an electoral system the fairness of its outputs: a system which defeats the causal equity of voters by privileging some or impeding others may yield results that go against the majority (when one exists). The distinction between high- and low-fidelity electoral systems could serve to more coherently express an intuition often shared by those who feel that their vote has been wasted: namely the intuition that, given the particular circumstances of the voting game, it is impossible, or at least very hard, to find a move or even a strategy which I (and other voters with my preference set) might reasonably make – though a candidate corresponding to my preferences could, or perhaps even does, exist. Many people who vote in systems of low fidelity will have shared this feeling.