Why Arrow's Theorem Matters for Political Theory—Even If Preference Cycles Never Occur

Sean Ingham

November 29, 2017

William Riker famously argued that Kenneth Arrow's impossibility theorem undermined the logical foundations of "populism," the view that in a democracy, laws and policies ought to express "the will of the people" (Riker, 1982).[1] In response, his critics have questioned the use of Arrow's theorem on the grounds that not all configurations of preferences are likely to occur in practice. The critics allege, in particular, that majority preference cycles, whose possibility the theorem exploits, are rare (Mackie, 2003; Regenwetter et al., 2006). In this essay, I argue that the critics' rejoinder to Riker misses the mark even if its factual claim about preferences is correct: Arrow's theorem and related results threaten the populist's principle of democratic legitimacy even if majority preference cycles never occur. In this particular context, the assumption of an 'unrestricted domain' is justified irrespective of the preferences citizens are likely to have.

The argument for this claim, aside from making a novel contribution to the debate about social choice theory and populism, also underscores the importance of a methodological rule for debating the assumptions of Arrow's theorem. One must distinguish between Arrow's theorem and its applications, and one should recognize that judgments about the plausibility of the theorem's assumptions are necessarily application-specific. In some applications, the justification for the unrestricted domain assumption stands or falls with facts about the kinds of preferences voters are likely to have. In others, like the theorem's application to principles of democratic legitimacy, facts about real-world distributions of preferences do not matter.

[1] I'm grateful to Wulf Gaertner, Lucas Stanczyk, and David Wiens for their comments on earlier drafts.

The theorem versus its applications

Arrow presented his theorem as a statement about "social welfare functions," by which he meant "a process or rule which, for each set of individual orderings $R_1, \ldots, R_n$ for alternative social states, ... states a corresponding social ordering of alternative social states, R" (Arrow, 1963, p. 23). Other commentators identify the theorem more narrowly with a statement about voting rules or judgments about social welfare. The relationship between these two interpretations—the social welfare function as a description of how society makes collective decisions or as a description of how an individual might make judgments of social welfare—was one of the early controversies Arrow addressed in the monograph's second edition.[2] In either case, however, the theorem is said to be a statement about the "aggregation" of individual preferences into a "social ordering" (which may represent the output of a voting rule or an individual's judgments about social welfare).

In my view, discussions of the theorem's applications would benefit from a different naming convention. As I will use the term here, *Arrow's impossibility theorem* names the statement following the definitional preamble:

> Let X and N be finite sets, with $|X| \geq 3$, and $|N| = n \geq 2$. Let $\mathcal{B}$ denote the set of complete binary relations on X, and let $\mathcal{R}$ denote the set of complete, transitive binary relations on X. Let ρ and ρ' denote generic n-tuples ("profiles") in $\mathcal{R}^n$ and let R_i and R'_i denote the ith components of the profiles with matching superscripts. For $R \in \mathcal{R}$, let P denote its asymmetric part: $(x, y) \in P$ if and only if $(x, y) \in R$ and $(y, x) \notin R$.
>
> *Arrow's impossibility theorem*: There is no function $f : \mathcal{R}^n \to \mathcal{B}$ such that
>
> (i) for all $\rho \in \mathcal{R}^n$, $f(\rho)$ is transitive;
>
> (ii) for all $x, y \in X$ and all $\rho \in \mathcal{R}^n$, if $(x, y) \in P_i$ for all $i \in N$, then $(y, x) \notin f(\rho)$;
>
> (iii) for all $x, y \in X$ and all $\rho, \rho' \in \mathcal{R}^n$, if
>
> $$(x, y) \in R_i \iff (x, y) \in R'_i, \quad \forall i \in N$$
>
> then $(x, y) \in f(\rho)$ if and only if $(x, y) \in f(\rho')$; and
>
> (iv) there is no $i \in N$ such that for all $\rho \in \mathcal{R}^n$ and $x, y \in X$, xP_iy implies $(y, x) \notin f(\rho)$.

[2] See section III 'What is the problem of social choice?' in Arrow's second edition of the monograph from 1963.

Under this naming convention, Arrow's impossibility theorem does not mention preferences, decisions, social welfare, or any other substantively meaningful concepts. Why adopt the convention of letting *Arrow's impossibility theorem* designate this colorless, formal statement about functions and binary relations, instead of a more eye-catching statement about voting rules or preference aggregation rules? We should do so because this abstract statement about binary relations, being more abstract, has many more potential applications than any of the less abstract statements about voting rules or preference aggregation rules, and we need a name for it. Moreover, we need to be able to distinguish between the different substantive applications, which, in turn, requires distinguishing between the abstract statement and its applications.

An "application" of the theorem is a statement identical to the formal theorem except that the abstract objects—X, $\mathcal{B}$ and $\mathcal{R}$, and $f : \mathcal{R}^n \to \mathcal{B}$—are all identified with the substantively meaningful objects or phenomena that they "represent" or "model." For a familiar example, let X be a set of candidates for office. Let $\mathcal{R}$ be the set of permissible ballots a voter can submit, where a ballot is a ranking (permitting ties) of the candidates; $\mathcal{R}^n$ is the set of possible combinations of ballots, one for each of n voters in the electorate. Assume a *democratic voting rule* generates a set of pairwise comparisons (again permitting ties) of the candidates for each combination of ballots; it has the same structure as a function $f : \mathcal{R}^n \to \mathcal{B}$. Then, applying the theorem, we can conclude that no democratic voting rule satisfies the transitivity condition (i), the unanimity condition (ii), the independence of irrelevant alternatives condition (iii), and the non-dictatorship condition (iv). An application of the theorem, like this one, associates its abstract mathematical objects with real-world objects or phenomena that have the same structure, or which can at least be treated as though they had the same structure for the purposes of inquiry. One gets different applications of the theorem by noticing all of the myriad examples in one's domain of inquiry of objects and phenomena exhibiting these same abstract structures. Other applications of Arrow's theorem, besides those I discuss here, include its applications to normative measurements of inequality and poverty (Sen, 1973, 1976); to judgments of similarity among counterfactual worlds in metaphysics (Morreau, 2010); and to criteria for scientific theory-selection (Okasha, 2011), among others.

Here is a second application, sometimes run together with the voting rule application by commentators who do not practice the rules of conceptual housekeeping I'm advocating. Let X again be a set of candidates, and let $\mathcal{R}$ be the set of possible complete, transitive preference rankings of the

candidates, so that $\mathcal{R}^n$ is the set of all possible combinations of preference rankings for n voters. Assume that *the will of the people* takes the form of a complete ranking of the candidates, and that one can define it as a function of voters' preferences over the candidates. In other words, any candidate definition of the concept can be represented by a function $f : \mathcal{R}^n \to \mathcal{B}$, which indicates how individuals' preferences determine the content of the popular will.[3] Then, applying the theorem, one can conclude that the definition, viewed as such a function, violates either condition (i), (ii), (iii), or (iv).

A democratic voting rule may have the same mathematical structure as a definition of 'the popular will', but the two concepts are nevertheless distinct, and any productive discussion of either application of the theorem requires keeping them distinct. Compare these statements about an election in which two parties, L and R, compete:

(a) The people prefer L to R.

(b) The election is decided by a democratic voting rule, and the rule ranks L above R after the votes are counted.

These two statements do not have the same meaning. Many democratic theorists believe in the existence of democratic voting rules but reject talk of a popular will. And among those who consider such talk meaningful, a certain kind of populist is prepared to make claims about the will of the people which he takes to be true independently of the results of the democratic process. As Jan-Werner Müller observes, populists often depict the people as "a homogeneous and morally unified body whose alleged will can be played off against actual election results in democracies" (Müller, 2016, p. 27). Such a populist claims there are facts about the popular will that are not just facts about the results of democratic voting. So, as many people use these terms, neither statement is equivalent to the other, even if the statements share abstract logical properties in virtue of which Arrow's theorem applies to each.

Below, I consider the theorem's application to statements about 'the will of the people' and the populist's principle of democratic legitimacy, according to which policies are legitimate only if they express the will of the people. In the context of that application, I argue that the assumption of an unrestricted domain (built into the definition of the function f) is justified irrespective of the preferences citizens are likely to have. This claim and in general any claim about the plausibility of the theorem's assumptions are

[3]This characterization of definitions precludes those under which, in some circumstances, the people lack a will. I return to this point below.

necessarily application-specific. This point is immediate once we agree to let *Arrow's theorem* refer to an abstract, mathematical theorem, instead of a substantive statement about voting rules or preference aggregation. It would be nonsensical to question the mathematical theorem's assumption that the domain of f is the unrestricted domain. Doing so would be like questioning the "assumption" that the Pythagorean theorem is about right triangles. These are not so much assumptions, capable of being false, as stipulated definitions of the abstract objects whose properties the theorems are describing. What one can intelligibly criticize or defend are assumptions that voting rules, the concept of the popular will, or other phenomena are well described by functions with the same structure as a function $f : \mathcal{R}^n \to \mathcal{B}$. In other words, one can question the assumptions of Arrow's theorem only in the context of one of its particular applications. In doing so, one will have to tailor one's arguments to the application at hand: in the context of one application, the assumption of an unrestricted domain may be justified for reasons altogether different from the reasons justifying it in the context of another application, and in other contexts it might not be justified at all. Indeed, I will argue that in the theorem's application to the populist's principle of legitimacy, the assumption of an unrestricted domain is justified irrespective of the preferences citizens are likely to have, even if, in other applications, its justification turns on these empirical facts.

Democratic legitimacy and the will of the people

The populist claims that

> (P1) the outcome of an election is legitimate only if it expresses the will of the people.

I present below what I think is the strongest argument for rejecting the populist's principle of legitimacy on the basis of Arrow's theorem and related results in social choice theory. This argument may resemble Riker's (1982) critique of populism at points, but one of several important differences is that the argument I present does not depend on assumptions about the occurrence of majority preference cycles, whereas Riker defends his conclusion in part by producing alleged historical instances of cycles.[4] Before presenting the argument, let me first refine the populist's claim.

[4]My argument also makes no claims about the manipulability of voting or the frequency with which different voting rules produce different outcomes from the same preferences; see Dowding (2006) for a discussion of these elements of Riker's argument.

Translated into a different idiom, (P1) is the claim that

> (P2) an election outcome is legitimate only if the people consider it no worse than any other feasible outcome.

I assume (P2) is an acceptable reformulation of (P1). It would be strange to describe an outcome as an expression of an agent's will if the agent considered it worse than a different feasible outcome.

One further refinement is in order. When the populist advances a claim like (P2), they are not merely making a statement about the world as it actually is. They are not merely saying, of the actual world, that the legitimate election outcomes happen to coincide with those which are best according to the people's preferences. They are instead making a claim about the necessary and sufficient conditions for legitimacy across a range of possible counterfactual circumstances. Anyone who affirms (P2) also means to endorse at least some judgments of the form 'if, counterfactually, the people preferred a candidate A to a candidate B, then the election of B would be illegitimate'. Let W denote the set of all possible worlds. If the populist's claim about legitimacy is to support any counterfactual judgments of this form, then its domain of application must be some subset $D \subset W$ of possible worlds that includes more than just the actual world. So (P2) can be restated as the following principle, where D includes possible worlds besides the actual one:

> (P3) In every possible world $w \in D$, an election outcome is legitimate only if the people consider it no worse than any other feasible outcome.

For now, let us suppose that $D = W$, that is, the relevant domain is the entire set of possible worlds. I revisit this assumption below.

Cyclical majorities as a thought experiment

What could it mean to say that the people prefer one outcome to another? One natural answer is that the people prefer one outcome, x, to another, y, if and only if a majority prefer x to y. I will refer to this as the *majoritarian interpretation* of the populist's principle of legitimacy.

Now imagine the following hypothetical scenario, which appears to be an effective counterexample to the populist's claim under the majoritarian interpretation.

> *Cyclical majorities.* There are three candidates, A, B, and C, competing in an election. All eligible voters participate; there

> is no electoral fraud; the candidates espouse reasonable and morally decent political views and run honest and fair campaigns; voters and candidates engage in sincere and high-quality deliberation; voters are well-informed about the candidates, hold reasonable and morally decent political views, and vote according to their sincere preferences; the background social and economic conditions leave no grounds for worrying about political inequality or other pathologies that might be thought to undermine the political legitimacy of an election. The Borda count is used to decide the election, and B is declared the winner after the ballots have been (accurately) tallied.
>
> But a majority prefer A to B, while another majority prefer B to C and another majority prefer C to A.

In *Cyclical majorities*, the election outcome is illegitimate according to the majoritarian interpretation of (P3), the populist's principle of legitimacy: it is false that the people consider candidate B no worse than any other outcome, under the majoritarian interpretation of the concept of the popular will. Moreover, *none* of the three possible election outcomes is legitimate, because for each possible outcome, a majority prefer an alternative outcome to it. Note that the populist's principle of legitimacy, (P3), entails these judgments even in variations of the hypothetical scenario in which different voting rules are used or in which voters vote differently and a different candidate wins. So long as one holds fixed the assumptions about citizens' preferences over the three candidates and adopts the majoritarian interpretation of the popular will, the principle implies that the election cannot have a legitimate outcome.

With a slight modification, the example also discredits the populist's principle when it is given an 'epistemic' interpretation according to which the people prefer an outcome x to an outcome y if and only if a majority of citizens sincerely judge that x is more congruent with an objectively defined common good than y. Some commentators have suggested that the epistemic conception avoids the difficulties studied in social choice theory (Cohen, 1986). But the same cycle in majorities' comparative jugdgments can arise; the source of difficulty is the aggregation of *rankings*, not preferences per se.

The judgment that none of the possible election outcomes is legitimate in *Cyclical majorities* should strike one as counterintuitive unless one is already firmly in the grip of the populist's theory. Intuitively speaking, if an election outcome is illegitimate, then there must be some actor or group of

actors whose actions or beliefs or preferences fail to comply with democratic norms and values, or some institution that violates the requisite democratic norms. If everyone has "democratically respectable" views and attitudes and acts accordingly, and the institutions in place comply with the relevant democratic principles, then it must be true of at least one of the feasible outcomes of the election that it would be legitimate if it were the outcome. But, in *Cyclical majorities*, none of the feasible outcomes satisfies the condition that the populist lays down as necessary for legitimacy, even though, by hypothesis, everyone has reasonable and morally decent political views, and the institutions and voting rules in place satisfy the relevant norms. Recall, again, that one can modify the example—swapping out the Borda count for some other voting rule or changing the assumptions about how people behave—and so long as one maintains the stipulations about citizens' preferences, the populist's principle entails the counterintuitive result that none of the outcomes is legitimate.

The example is meant to show that the populist's principle does a poor job of capturing intuitions about democratic legitimacy. The explanation for the failure is straightforward. Under the majoritarian interpretation of the popular will, there are circumstances in which no outcome is "best" according to the popular will. These circumstances, defined by a certain kind of configuration of citizens' preferences, can obtain even if one stipulates idealized citizens, parties, institutions, background conditions, etc. The example suggests that our intuitions about democratic legitimacy track facts about the democratic institutions in place and whether people comply with certain democratic norms, not facts about whether the outcome of the democratic process can be intelligibly viewed as an expression of a popular will.[5] The prefatory information about the hypothetical election—the stipulation that all the actors are playing by the rules and living up to various democratic ideals—is meant to elicit these intuitions.

Might the populist reply that each outcome is legitimate in *Cyclical majorities*? After all, each is, in a sense, "on a par" with respect to majorities' preferences, so perhaps the populist should say that the people are indifferent, judging each as good as any other. This reply involves a departure from the majoritarian interpretation of the popular will, as defined above, and *Cyclical majorities* is only meant to be a counterexample to the populist's principle under that interpretation. I consider below whether the populist

[5]For a theory of democratic legitimacy that engages with the challenges from social choice theory and identifies legitimacy with properties of the decision-making process, see Patty and Penn (2014).

can save (P3) under alternative interpretations of the popular will.

It would be no defense of the populist's principle to point out that the hypothetical example is "unrealistic" or improbable. Hypothetical scenarios like this one are meant to be "intuition pumps" (Dennett, 2013). In order to elicit and clarify the intuitions that conflict with a principle, like the populist's, one contemplates a scenario that throws its critical features into sharp relief. One may have to contemplate a rather far-fetched scenario in order to isolate certain implications of the principle. But that is fine, because the scenario is merely a tool to help one think through the principle's implications. It is not intended as a prediction of what might happen. Constructing improbable hypothetical scenarios as intuition pumps is a widely accepted method of argument in political theory and philosophy. Think of the famous trolley problem in ethics, and how variations on it, all far-fetched, are used as counterexamples to utilitarianism and other proposed principles of ethics. The method is familiar in democratic theory, too. "Let us transport ourselves into a hypothetical country that, in a democratic way, practices the persecution of Christians, the burning of witches, and the slaughtering of Jews," Schumpeter writes, offering the scenario as a counterexample to the idea that the results of democratic procedures always merit respect (Schumpeter, 1942, p. 242). Brennan (2016, p. 12) uses a thought experiment in the same vein to discredit proceduralist theories of democracy.

In political theory and philosophy, some writers dispute the relevance of outlandish hypothetical scenarios. But the scenarios responsible for controversy are truly *outlandish*, not merely far-fetched. For example, as part of a critique of the principle of self-ownership, Kaspar Lippert-Rasmussen (2008) asks his reader to imagine "people [who] are born with huge bodies they can barely move, bodies with two hundred legs and arms." Faced with an "other-worldly" scenario such as this one, a critic might contend that the principle which the scenario is supposed to discredit (the principle of self-ownership here) is only meant to apply under special circumstances, which do not obtain in the hypothetical scenario. Or the critic might argue that the scenario is so outlandish that it is hard to imagine—hard to imagine all of the implications of its stipulations—and doubtful whether one's intuitions about the scenario are worth taking seriously (Elster, 2011). Note, however, that neither of these responses is an objection to the hypothetical scenario solely on the grounds that it is improbable. The mere improbability of a scenario is neither a reason to deny it belongs to a principle's domain of application nor a reason to question the validity of our intuitive judgments about it.

Might a defender of the populist's principle claim that *Cyclical majori-*

ties is not merely improbable but also outlandish, arguing either that a democratic principle of legitimacy is only meant to apply under "realistic" conditions that are not met in the scenario, or that the scenario is outlandish enough that we should not put stock in our intuitions about it? Take the latter argument first. What (if anything) makes a hypothetical scenario objectionably outlandish is open to reasonable debate, but there is no need to take a firm position on this question. It suffices to show that the stipulation of majority preference cycles resembles in relevant respects other stipulations which we agree do not render hypothetical scenarios objectionably outlandish. To this end, consider the stipulation of universal compliance with democratic norms in *Cyclical majorities*. I assume the reader agrees that this stipulation does not render the scenario objectionably outlandish, even though it makes the scenario extremely improbable. Why not? We are familiar with *actual* individuals who comply with democratic norms and values, hold reasonable and morally decent political views, and so on. In imagining a hypothetical scenario in which all citizens fit this description, we are imagining a scenario which is admittedly improbable in the extreme but which could nevertheless arise in the world as we know it, by causal processes similar to those producing the admirable individuals we are familiar with in actual life. This familiarity arguably explains why we are justified in trusting our ability to accurately imagine and assess a hypothetical scenario in which everyone's character and values have been shaped by these same causal processes.

But in this respect, the stipulation of cyclical majorities is no different from the stipulation of good democratic citizenship. Let us fill in *Cyclical majorities* with a bit more detail. Suppose that candidate A is on the left, candidate C is on the right, and candidate B is a moderate. The population splits into three camps: one third prefer A to B to C, another third prefer B to C to A, and another third prefer C to A to B. The final group may seem odd in viewing the moderate as worse than both the left- and the right-wing candidates. Their preferences fail to be single-peaked with respect to the left-right ordering. But scholars find in the data *actual* individuals who are eccentric in this way. Regenwetter et al. (2006) conclude from their empirical study of actual preference distributions that, unsurprisingly, the assumption that *everyone*'s preferences are single-peaked never holds in reality. The assumption of single-peaked preferences may be a reasonable *approximation* of actual electorates, but it is only an approximation. A scenario in which a full third of the electorate holding preferences that fail to be single-peaked with respect to a left-right ordering may be improbable, but it is a scenario which could, despite its improbability, arise in the world

as we know it, by causal processes similar to those producing the eccentric individuals we observe in the actual world. It is no harder to imagine than an electorate whose members all live up to democratic norms and values. These observations, I submit, are sufficient reason to conclude that *Cyclical majorities* is not objectionably outlandish, however improbable it may be. We are justified in trusting our ability to accurately imagine and assess the scenario described.

Consider next the contention that a principle of democratic legitimacy is only intended to characterize the necessary and sufficient conditions for legitimacy across a special subset $D \subset W$ of possible worlds that excludes scenarios like *Cyclical majorities*. Absent reasons for delineating the domain of relevant possible worlds just so, the domain restriction looks like an ad hoc attempt to save a principle from the counterexample. What might these reasons be? The basis for excluding *Cyclical majorities*, or any other scenario, cannot be mere probability of occurrence. What reason could there be then? The essential feature of the hypothetical scenario is a certain diversity of perspectives and disagreement: citizens not only disagree about which candidate is best, but, in a sense, disagree about whether their location on a left-right continuum is even relevant. That, anyway, is one possible explanation for why some citizens' preferences fail to be single-peaked with respect to this scale (Dryzek and List, 2003; Knight and Johnson, 2011). But disagreement is part of the circumstances of politics, and it is precisely under conditions of diversity and disagreement that democracy is thought to have its unique claim to legitimacy (Bellamy, 2007; Waldron, 1999). Presented with circumstances in which all citizens and political actors are committed to democratic norms and values, it would be strange to deny that principles of democratic legitimacy apply simply because citizens' preferences over candidates exhibit too much heterogeneity.

Here is a second argument for identifying the domain of principles of democratic legitimacy with Arrow's unrestricted domain of preference profiles. Assume that the actual world belongs to the relevant domain, and recall the point, from above, that one finds all manner of preferences in actual electorates, even if some preference rankings are only found among small numbers of voters. That is to say: for any preference relation $R \in \mathcal{R}$ defined over a small set of candidates, one can find, in a large electorate, an actual voter whose preferences are R. Now imagine a process of deliberation that affects the number of voters holding each of the preference rankings observed in the actual world. Any possible configuration of preferences is the potential result of such a process because, by hypothesis, the actual electorate contains every possible preference relation. There is, I

submit, no good reason for holding that principles of democratic legitimacy apply to the actual world but not the worlds resulting from these counterfactual processes of deliberation.[6] Thus, the relevant domain for principles of democratic legitimacy must be the unrestricted domain, as far as citizens' preferences go.

If the populist has no grounds for excluding *Cyclical majorities* from the domain in which principles of democratic legitimacy are supposed to apply, and no reasons for discounting the intuitions about the scenario when they conflict with the populist's principle, what are the remaining replies to the counterexample? The populist might concede that *Cyclical majorities* is a counterexample to the principle of legitimacy but only when the principle is given its majoritarian interpretation. Perhaps there is another plausible definition of the concept of the popular will relative to which the populist's principle would be protected from counterexamples like this one. I address this possibility next.

Arrow's theorem and related results

Arrow's impossibility theorem can help us think about the plausibility of this response. We can identify candidate definitions of the concept of the popular will with the functions that Arrow's theorem describes. To follow the argument, first banish from the mind the application of Arrow's theorem to voting rules. The argument does not depend on any assumptions about the voting rule or other institutions in place, how citizens vote or otherwise behave, or the preferences anyone has. It concerns the possible definitions of a concept.

The first assumption is that one can associate any candidate definition of the concept of the popular will with a function $f : \mathcal{R}^n \to \mathcal{B}$, which describes how alternative election outcomes compare according to the popular will, as a function of citizens' preferences (holding fixed other features of possible worlds that also bear on such comparisons, if there are any).[7] For example, the majoritarian definition, from above, runs: for any two alternatives x and y, and any possible world w, the people weakly prefer an alternative x to an alternative y at w if and only if a majority weakly prefer x to y at w. This

[6] Some of these counterfactual processes are highly improbable, but we have already ruled out probability as a criterion for determining relevance.

[7] One should get off the train at this early juncture if one thinks facts about the popular will depend on more than just citizens' preferences. As is well-known in the context of social welfare functions, Arrow's impossibility result can be avoided if the aggregation function has a richer informational base (Sen, 1977).

definition corresponds to the function $f_m : \mathcal{R}^n \to \mathcal{B}$ such that $(x, y) \in f_m(\rho)$ if and only if $|\{L \subset N | xR_l y, \forall l \in L\}| > n/2$. Any alternative definition of the concept will be associated with its own function $f : \mathcal{R}^n \to \mathcal{B}$, indicating how to make judgments about the people's preferences as a function of individuals' preferences.

I argued above that for every profile of preferences there is a possible world which is in the relevant domain of the populist's principle of legitimacy and in which citizens have those preferences. The populist's principle specifies well-defined conditions for legitimacy across its domain only if the concept of the popular will is well-defined across this domain. That is why the definition of the concept must be associated with a function defined on the unrestricted domain $\mathcal{R}^n$.

This assumption rules out the possibility of a definition under which the people sometimes lack a will or have incomplete preferences, although it permits the scenario of universal indifference among the candidates. But allowing for this possibility is unlikely to help the populist who claims that an election outcome is legitimate only if it expresses the popular will. One could reasonably say of scenarios like *Cyclical majorities* that the people lack a will, but then one cannot say that any of the election outcomes express the popular will. So the populist's principle will still entail that none of the possible outcomes is legitimate, contrary to the intuition that stipulations about the scenario are meant to elicit.

In this context, Arrow's conditions can be viewed as proposed constraints on adequate definitions of the concept of the popular will. On any plausible definition, we will want to say that the people prefer x to y if every citizen strictly prefers x to y. So f will need to satisfy condition (ii), the weak Pareto condition. And, arguably, our judgments about which of two alternatives the people prefer should be sensitive only to citizens' preferences over those two alternatives. Saying that the people prefer x to y is, after all, supposed to be a means of expressing information about citizens' attitudes *towards* x *and* y, not their attitudes towards other pairs of alternatives. So f will need to satisfy condition (iii), independence of irrelevant alternatives.[8] And, least controversially, on any adequate definition of the concept, there is no one single individual whose preferences alone determine the content of the popular will. So f will need to satisfy the condition (iv), the non-dictatorship condition. Then, applying Arrow's theorem, we know that f

[8] For an interesting critique of independence and an argument for why the Borda count represents a plausible interpretation of the popular will, see Saari (2003), especially pp. 342–349.

will violate transitivity. Thus, on any adequate definition of the concept of the popular will, there are possible worlds in which the people weakly prefer an alternative x to y, also weakly prefer y to z, but fail to weakly prefer x to z.

Some might wish to push this argument further and argue that transitivity is itself a constraint on any adequate definition of the concept of the popular will. If it is, then Arrow's theorem implies that the concept does not permit an adequate definition. This move is dubious, however; the concept of intransitive preferences, whether imputed to an individual or a group, is strange but not oxymoronic.

Intransitivity in the popular will is compatible with there being a legitimate outcome according to the populist's principle of legitimacy, as it is expressed in (P3). That principle merely requires the existence of an alternative that the people consider no worse than any other, and intransitivity does not rule out the possibility of such an alternative.[9] In light of this observation, the populist might deny that Arrow's theorem poses any challenge to his account of democratic legitimacy. In *Cyclical majorities*, none of the feasible outcomes satisfies the necessary condition for legitimacy according to the majoritarian interpretation of the populist's principle, and that fact was a reason to reject the populist's principle of legitimacy, provided one has the intuitions that the thought experiment tries to elicit. When we entertain alternatives to the majoritarian interpretation of the popular will, Arrow's theorem tells us that every definition sometimes implies intransitive judgments about which alternatives are better and worse according to the people, unless it violates one of the other conditions for adequacy. But, the populist might respond, this conclusion does not imperil the populist's principle in the same manner *Cyclical majorities* threatened to, because there may still be an adequate definition which always allows one to identify a "best" alternative—an outcome that the people judge no worse than any other—even though it entails intransitive judgments. Transitivity is not necessary for the existence of a best alternative. If there is such a definition, then, using this definition of the popular will, the necessary condition for legitimacy can always be satisfied, and there will be no counterexample analogous to *Cyclical majorities*.

Arrow's theorem may not tell us that such a definition is impossible, but a host of related results, inspired by Arrow's contribution, arguably do. The results I have in mind in effect show that any definition of the concept of

[9] See the discussion of the maximal set and conditions for its nonemptiness in Austin-Smith and Banks (1999, ch. 1).

the popular will entails either a mapping $f : \mathcal{R}^n \to \mathcal{B}$ that violates plausible constraints on adequate definitions or the existence of scenarios that are analogous to *Cyclical majorities*. Using these results, the following template generates counterexamples to different versions of the populist's principle of legitimacy, depending on how the concept of the popular will is defined. Suppose the populist has supplied a definition, and let $f : \mathcal{R}^n \to \mathcal{B}$ be the mapping entailed by the definition, whatever it is.

> *Counterexample template.* There are at least k candidates for office. All eligible voters participate; there is no electoral fraud; the candidates espouse reasonable and morally decent political views and run honest and fair campaigns; voters and candidates engage in sincere and high-quality deliberation; voters are well-informed about the candidates, hold reasonable and morally decent political views, and vote according to their sincere preferences; the background social and economic conditions leave no grounds for worrying about political inequality or other pathologies that might be thought to undermine the political legitimacy of an election. The Borda count is used to decide the election.
>
> Yet citizens' preferences over the candidates are such that for every candidate, one of other feasible candidates is better according to the people, i.e., according to the ranking generated by f from citizens' preferences.

The template is not meant to generate a counterexample to the populist's principle under any definition of the concept of the popular will. For some definitions, the template fails to produce valid counterexamples, because for some choices of f, the hypothesized scenario is logically impossible. For example, if f is a mapping, like the Borda count, which generates an acyclic ranking for every profile of preferences, then the scenario is not a well-defined possibility, because there is guaranteed to be a "best" candidate relative to an acyclic ranking of a finite number of candidates. But we already know that the template works for at least one possible definition: when f is "majority rule" and $k = 3$, we have the special case of *Cyclical majorities*.

What more can we say about other possible definitions, besides those corresponding to the Borda count and majority rule? Quite a bit, it turns out. For a large class of functions f, there is a critical number k such that f sometimes produces cyclical rankings when there are at least k alternatives (Banks, 1995; Austin-Smith and Banks, 1999). Here is one result, for the

sake of illustration: if f is decisive, neutral, monotonic, and anonymous, then it is a "q-rule," meaning it ranks one outcome x above another y whenever at least q citizens strictly prefer x to y (Austin-Smith and Banks, 1999, theorem 3.7). Any q-rule with $q < n$ will generate a cyclical ranking for some profile of preferences if $k \geq 1/(1-q/n)$ (Austin-Smith and Banks, 1999, corollary 3.1). And if such a profile exists, then there also exists a profile at which every alternative is judged worse than another alternative according to q citizens. (One can get the second profile from the first by moving all alternatives not contained in the cycle down in each person's preference ranking.) One can therefore construct a scenario in which there are $k \geq 1/(1-q/n)$ candidates, citizens have these preferences over the candidates, and all of the other facts from *Counterexample template* obtain. For example, if our definition says that an outcome x is better than an outcome y whenever at least three-fourths of voters judge x to be better than y (i.e., $q = 3n/4$), then we can use the template to construct a hypothetical scenario in which there are four candidates and every candidate is worse than another candidate according to the popular will. In this scenario, none of the four possible outcomes will satisfy the populist's necessary condition for legitimacy.

And yet surely it must be possible for the election described in the template to have a legitimate outcome, given that everyone is an exemplary democrat, complying with whatever democratic norms and values citizens and candidates should comply with. Not so, according to the populist's principle of legitimacy. At least, that is the principle's implausible verdict if judgments about the people's preferences are a function of citizens' preferences, and the function satisfies the conditions of decisiveness, neutrality, monotonicity, and anonymity. If these conditions are defensible as constraints on adequate definitions of the concept of the popular will, then the principle fails to accommodate our intuitions about legitimacy in the scenario described in the template. That is a reason to reject the populist's principle of legitimacy.

Naturally, some people will question whether conditions like neutrality and monotonicity express reasonable constraints, just as some question whether independence of irrelevant alternatives is a reasonable constraint. While space does not permit a full discussion of the issue, let me make one pertinent observation. These conditions may be justified in the context of one application but not another, and we must guard against conflating different applications. Let me illustrate the point with Arrow's independence condition. The relevant question here, in the context of the argument I have made about the populist's principle of legitimacy, is not whether a reasonable method of making decisions would allow a choice between two

alternatives to be influenced by individuals' preferences over other alternatives. The relevant question is rather the following. If a definition of the popular will entails that facts about the people's preferences over two alternatives x and y can change even without any change in any individual voter's preferences over x and y, is the definition plausibly viewed as an attempt to clarify the intuitive notion of the popular will? Or has it then strayed so far from ordinary intuitions about the concept of the popular will that calling it a definition of this concept is false advertising? We must abide by the methodological rule advocated above and pay attention to the particular context, and what the function f is representing, when we evaluate the independence assumption. Reasons to reject the assumption in some contexts are not necessarily reasons to reject it as a constraint on adequate definitions of the concept of the popular will.

These methodological strictures do not mean that formal arguments, which abstract from any substantive application, have no place in justifications of Arrow's conditions. For example, in defense of the independence condition, scholars have leveraged formal results that abstract from particular applications and show how independence can be replaced with weaker or more intuitively appealing conditions. For example, Patty and Penn (2014, theorem 3) show that independence is equivalent to a more intuitive condition they call "unilateral flip independence". Such results help us understand the meaning and formal implications of the independence condition, and against some critiques of the assumption, these results are an effective rebuttal. All of this is consistent with the methodological point I am emphasizing: in the final analysis, our judgments about the plausibility and reasonableness of an assumption like independence are judgments about the application of formal definitions in this or that substantive application, and we should expect these judgments to be sensitive to the context, even though formal, application-neutral results may also bear on our judgments.

The debate about the unrestricted domain assumption illustrates this methodological point as well. I have argued that the populist's principle of legitimacy should be judged by its implications across a range of possible scenarios that comprises Arrow's unrestricted domain. More precisely: for every profile of preferences, there is a possible world which is in the domain of principles of democratic legitimacy and in which citizens have those preferences. The scenario described in *Cyclical majorities* and the other scenarios constructed from *Counterexample template* may be improbable, but that is no reason to dismiss their relevance. Their improbability is neither a reason to discount our intuitive judgments about legitimacy in these hypothetical scenarios nor a reason to exclude them from the domain of possible worlds

in which we expect a principle of democratic legitimacy to yield intuitively acceptable judgments.

That argument was a justification of Arrow's assumption of an unrestricted domain *in the context of this particular application of the theorem.* In the context of other applications of Arrow's theorem, the assumption of an unrestricted domain might be unjustified. Scholars have often suggested that the reason to worry about majority preference cycles is that majority voting would produce undesirable instability and "chaos" if there are cycles in majorities' preferences. This worry, and any other worry about the *effects* of democratic voting rules when majorities have cyclical preferences, should presumably be scaled to the probability of such effects. The worry might be put to rest if, as scholars have argued, the institutional "structures" found in democracies induce stability in democratic decision-making, even in the presence of majority preference cycles (Shepsle and Weingast, 1981). Or it might be put to rest by empirical arguments that majority preference cycles are rare (Mackie, 2003; Regenwetter et al., 2006). In the context of this application, there could be good reason to care about the probability of majority preference cycles.

The populist's claim about democratic legitimacy, however, is not a claim about the effects of democratic voting rules. It is a conceptual claim about the necessary conditions for democratic legitimacy. A good thought experiment can embarrass the conceptual claim by showing that it has absurd implications in some hypothetical scenario, even if the scenario is improbable. These observations underscore the importance of distinguishing between different applications of Arrow's impossibility theorem, which, in turn, presupposes a distinction between the abstract, purely formal statement and its substantive applications.

Conclusion

Despite all the attention it has received, Arrow's impossibility theorem may still be underrated in political theory. Too many scholars still identify it with one of its applications, like its application to voting rules. Just as one would undervalue calculus if one identified it with the study of changes in the rate of motion of physical objects, one will undervalue Arrow's impossibility theorem if one identifies it with a statement about voting rules. To fully appreciate the many rich implications of the theorem, political theorists should first learn to see it as a dry, bloodless statement about binary relations and functions. We will then be in a good position to discover new substantively

interesting applications and shed new light on old controversies.

As an example of this general methodological point, I have explained how Arrow's theorem and related results pose a challenge to the populist's principle of democratic legitimacy irrespective of the preferences citizens are likely to have. Applied to possible definitions of the concept of the popular will, the theorem shows that on any adequate definition of the concept, the people's preferences will sometimes be intransitive. Related results arguably allow us to conclude that on any adequate definition, there will be scenarios in which the people consider every possible election outcome worse than another, and in these scenarios, no outcome will satisfy the populist's principle of legitimacy. The theorems' assumption of an unrestricted domain is justified in this context, but not because any configuration of preferences is equally or sufficiently likely. In other applications, its justification might depend on the preferences citizens are likely to have. But it is justified here because the domain of possible scenarios in which a principle of democratic legitimacy should yield intuitively acceptable answers includes all kinds of possible scenarios, including improbable ones.

References

Arrow, Kenneth. 1963. *Social Choice and Individual Values.* Second ed. New Haven, CT: Yale University Press.

Austin-Smith, David and Jeffrey Banks. 1999. *Positive Political Theory I: Collective Preference.* Ann Arbor, MI: University of Michigan Press.

Banks, Jeffrey S. 1995. "Acyclic Social Choice from Finite Sets." *Social Choice and Welfare* 12(3):293–310.

Bellamy, Richard. 2007. *Political Constitutionalism: A Republican Defence of the Constitutionality of Democracy.* Cambridge: Cambridge University Press.

Brennan, Jason. 2016. *Against Democracy.* Princeton, NJ: Princeton University Press.

Cohen, Joshua. 1986. "An epistemic conception of democracy." *Ethics* 97(1):26–38.

Dennett, Daniel C. 2013. *Intuition Pumps and Other Tools for Thinking.* WW Norton & Company.

Dowding, Keith. 2006. "Can populism be defended? William Riker, Gerry Mackie and the interpretation of democracy." *Government and Opposition* 41(3):327–346.

Dryzek, John S. and Christian List. 2003. "Social Choice Theory and Deliberative Democracy: A Reconciliation." *British Journal of Political Science* 33:1–28.

Elster, Jakob. 2011. "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28(3):241–258.

Knight, Jack and James Johnson. 2011. *The Priority of Democracy: Political Consequences of Pragmatism*. Princeton, NJ: Princeton University Press.

Lippert-Rasmussen, Kasper. 2008. "Against Self-Ownership: There Are No Fact-Insensitive Ownership Rights over One's Body." *Philosophy & Public Affairs* 36(1):86–118.

Mackie, Gerry. 2003. *Democracy Defended*. Cambridge: Cambridge University Press.

Morreau, Michael. 2010. "It Simply Does Not Add Up: Trouble with Overall Similarity." *The Journal of Philosophy* 107(9):469–490.

Müller, Jan-Werner. 2016. *What is Populism?* Philadelphia, PA: University of Pennsylvania Press.

Okasha, Samir. 2011. "Theory Choice and Social Choice: Kuhn versus Arrow." *Mind* 120:83–115.

Patty, John W. and Elizabeth Maggie Penn. 2014. *Social Choice and Legitimacy: The Possibilities of Impossibility*. Cambridge: Cambridge University Press.

Regenwetter, Michel, Bernard Grofman, A. A. J. Marley and Ilia Tsetlin. 2006. *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge: Cambridge University Press.

Riker, William. 1982. *Liberalism Against Populism*. New York: Waveland Press.

Saari, Donald G. 2003. "Capturing the "Will of the People"." *Ethics* 113(2):333–349.

Schumpeter, Joseph. 1942. *Capitalism, Socialism, and Democracy.* New York: Harper & Brothers.

Sen, Amartya. 1973. *On economic inequality.* Oxford: Oxford University Press.

Sen, Amartya. 1976. "Poverty: an ordinal approach to measurement." *Econometrica* pp. 219–231.

Sen, Amartya. 1977. "On weights and measures: informational constraints in social welfare analysis." *Econometrica* pp. 1539–1572.

Shepsle, Kenneth A. and Barry R. Weingast. 1981. "Structure-Induced Equilibrium and Legislative Choice." *Public Choice* 37(3):503–519.

Waldron, Jeremy. 1999. *Law and Disagreement.* Oxford: Oxford University Press.