A Formal Theory of Democratic Deliberation

Hun Chung [*] John Duggan [†]

February 17, 2019

Abstract

Inspired by impossibility theorems of social choice theory, many democratic theorists have argued that aggregative forms of democracy cannot lend full democratic justification for the collective decisions reached. Hence, democratic theorists have turned their attention to deliberative democracy, according to which "outcomes are democratically legitimate if and only if they could be the object of a free and reasoned agreement among equals." (Cohen 1997a: 73) However, relatively little work has been done to offer a formal theory of democratic deliberation. This paper helps fill that gap by offering a formal theory of three different modes of democratic deliberation: myopic discussion, constructive discussion, and debate. In either form of discussion, positions are considered according to an exogenous protocol and arguments applied to them, whereas in a debate, two participants who have diametrically opposed preferences take turns and propose positions with supporting reasons/arguments. We show that myopic discussion suffers from indeterminacy of long run outcomes, while constructive discussion and debate are conclusive, *i.e.,* both forms of deliberation converge to a position that is maximally justified according to at least one reason/argument. Finally, unlike the other two modes of deliberation, debate is path independent and converges to a unique compromise position, irrespective of the initial status quo.

Keywords Democratic Theory; Deliberative Democracy; Democratic Justification; Legitimacy; Formal Political Theory

[*]Associate Professor, School of Political Science and Economics, Waseda University
[†]Professor of Political Science and Economics, University of Rochester

1 Introduction: The Need for a Formal Theory of Democratic Deliberation

Democracy is an institutional arrangement for making binding collective decisions; specifically, it is concerned with making binding collective decisions among a wide range of people residing in a political body and who are all regarded as free and equal. This contrasts with other types of political systems – such as a dictatorship or an aristocracy – in which only a few enjoy such moral status. One may not always agree with the specific collective decision reached through a democratic process, but the implementation of a collective decision, once reached, involves the use of state force and political coercion. Hence, to respect the moral status of free and equal citizens (including those who disagree with the specific policy), a collective decision reached in a democracy must be *justified.* But, how?

Many scholars have pointed out that the mere fact that a collective decision has been reached through a particular voting procedure, say, a majority vote, is in itself insufficient to lend full justification or legitimacy to the collective decision at hand. It has been known since the work of Marquis de Condorcet in the 18th century that majority voting can produce *voting cycles,* even when each voter has a transitive ranking of alternatives. The existence of voting cycles causes problems for democratic justification or legitimacy, as they create the possibility that given any choice by society, there is a majority of citizens who prefer a different social alternative to the one chosen. This means that there exists no social alternative that can be unambiguously regarded as the "best" social choice, justifiably superior to all others.

Kenneth Arrow (1951/1963) extended Condorcet's insight and went further by showing that every voting mechanism will fail to satisfy at least one among a number of reasonable and seemingly innocuous conditions of fairness and rationality.[1] Kenneth May (1952) has shown that the only voting procedure that treats both the voters as well as the social alternatives impartially and responds positively to changes in voter preferences is majority rule. However, Charles Plott (1967) has shown that whenever there are multiple issue dimensions, the majority core (defined as the set of social alternatives that cannot be beaten by a pairwise majority vote) is generically empty. Richard McKelvey (1976,1979) and Norman Schofield (1978) have shown that in this situation, the top cycle engulfs the entire space of alternatives, so that a suitable choice of voting agenda can lead society to eventually adopt any given social alternative starting from any given status quo by a sequence of pairwise votes. The implication is that this makes it possible for those who have the power to control the agenda to obtain any desired outcome by strategically manipulating the agenda. A related idea is that voters themselves can ma-

[1] When there are at least three alternatives and the domain of preferences is unrestricted, the following axioms are inconsistent: Pareto Efficiency, Independence of Irrelevant Alternatives, No Dictatorship, and Social Rationality.

nipulate choices by misrepresenting their preferences; Allan Gibbard (1973) and Mark Satterthwaite (1975) have shown that the only single-valued social choice functions that are free of strategic manipulation are dictatorial, and John Duggan and Thomas Schwartz (2000) have shown that the impossibility result extends even when social ties are possible. This is just a representative handful of results published in the field of social choice theory. As one can see, the field of social choice theory is replete with impossibility theorems.

William Riker has argued that these negative results of social choice theory demonstrate that electoral outcomes are simply "meaningless" and can never be regarded as the "fair and true amalgamations of the voters' judgments." (Riker 1982: 238) As a consequence, elections can never truly justify a given collective decision, and the only meaningful role elections may perform is to periodically replace incompetent and disliked political officials to prevent society from falling into tyranny.[2] (Riker 1982: 239–246)

Those who wished to preserve the notions of democratic justification and legitimacy through aggregative voting mechanisms simply denied the practical relevance of social choice theory. (Mackie 2003) Others who thought aggregative voting mechanisms fall short of fully justifying collective decisions, but who, nonetheless, wished to preserve a notion of democratic justification or legitimacy in democratic theory, turned their theoretical attention to *deliberative democracy*. Deliberative democratic theory is founded on the basic principle that "outcomes are democratically legitimate if and only if they could be the object of a free and reasoned agreement among equals."(Cohen 1997a: 73) Amy Gutman and Dennis Thomson characterize this as the "reason-giving requirement" and explain that it is the "first and most important characteristic" of a deliberative democracy. (Gutman and Thomson 2004: 3) In other words, according to advocates of deliberative democracy, the fact that a given political outcome has survived the test of reasoned public deliberation serves as the basis for its very justification and legitimacy.

Then, how exactly does this process of reasoned public deliberation – that is, the process of presenting arguments and exchanging reasons for or against proposed options – confer justification for the proposals that survive this process?

Some scholars have argued that reasoned public deliberation lends justification because the proposals that are sustained and survive through the process of deliberation are simply *better* in terms of its overall quality. Simply put, outcomes of deliberative procedures tend to be more rational, better supported by hard or soft evidence, and can even be closer to some objective standard of correctness or truth. (Bohman and Rehg: xix) This way of explaining the value of public deliberation and its connection to political justification presumes that there exist some procedural-independent criteria of rightness or correctness that the procedure of public deliberation is able to track. Many epistemic democrats, such as David Estlund (1997) and Helene Landemore

[2]Riker calls this conception of democracy "liberal," as opposed to "populist."

(2013), hold this view.

For epistemic democrats, the basic aim of democracy is to "track the truth." (List and Goodin 2001: 277) In defending a view that he calls "epistemic proceduralism," Estlund explains that "democratic legitimacy requires" not only that "the procedure is procedurally fair" but also that it "can be held, in terms acceptable to all reasonable citizens, to be *epistemically the best* among those that are better than random." (Estlund 1997: 174 emphasis ours) Similarly, Landemore advocates what she calls "inclusive deliberation," in which reasoned public deliberation occurs among a large and cognitively diverse group of people in search for the political truth. (Landemore 2013: chapter 4) Informed by the "Diversity Trumps Ability Theorem" by Lu Hong and Scott Page (Hong and Page 2004: Theorem 1, 16388), Landemore claims that such an arrangement will more likely lead us to political outcomes that meet a set of procedural-independent standards of correctness than leaving public deliberation to a small number of experts. (Landemore 2013: 117, 210–211) In short, for epistemic democrats, public deliberation lends justification to its outcomes because post-deliberation outcomes are more likely to be objectively correct.[3]

Another group of scholars claim that post-deliberation outcomes are more justified than simple non-deliberative aggregative outcomes because the very procedure of reasoned public deliberation embodies or manifests core values of basic human morality and political justice, and it forces participants to be attentive toward the common good. For instance, Thomas Christiano argues that justice demands that citizens show mutual respect to each other, and "[i]n politics, expressing respect for persons who will be affected by a decision involves, in addition to giving them a vote in the decision, seeking out their views and engaging them in" public deliberation. (Christiano 1997: 252) By doing so, individuals will "have *equality* in the cognitive conditions of democratic decision making" which "is a requirement of justice." (Christiano 1997: 253) Similarly, Jack Knight and James Johnson claim that justice demands that citizens enjoy "*equal opportunity of access to political influence*" (Knight and Johnson 1997: 280) and post-deliberation outcomes are more justified to the extent that the very process of reasoned public deliberation embodies such a requirement of justice. John Rawls makes a similar point when he claims that "the principle of reciprocity" requires that "our exercise of political power is proper only when we sincerely believe that the reasons we offer for our political action may reasonably be accepted by other citizens as a justification of those actions." (Rawls 1997: 133–134) Joshua Cohen claims that post-deliberation outcomes are more justified because the very procedure of reasoned public deliberation itself embodies and tends to promote the common good and autonomy. (Cohen 1997a: 76–77)

[3]Sean Ingham (2012) casts doubts on justifying deliberative democracy on such epistemic grounds as doing so conflicts with what he calls the "non-convergence constraint" (according to which there could always remain rational disagreement in political matters) and what he calls the "constraint on evidence" (according to which justification for a democratic procedure should not presuppose substantive agreements on the outcomes of such procedures.)

Finally, a number of scholars have recently argued that reasoned public deliberation may also complement (or even nullify the need for) aggregative voting mechanisms by helping us escape the various impossibility results of social choice theory. According to Jon Elster's account of Jürgen Habermas's "ideal speech situation," the goal of politics is "[n]ot optimal compromise, but unanimous agreement," and after going through reasoned public deliberation "there would not be any need for aggregating mechanism, since a rational discussion would tend to produce unanimous preferences." (Elster 1997: 11-12) Other scholars have claimed that hoping to reach such an unanimous agreement in a diverse and pluralistic society is unrealistic. (Gaus 1997; Knight and Johnson 1994, 2007; Rawls 1997, 1999/2003) Instead, noting that the possibility of cycling and instability of aggregative voting mechanisms usually occur when there are multiple issue dimensions, Knight and Johnson explain that democratic deliberation may help us escape the many impossibility results of social choice theory by "induc[ing] a shared understanding regarding the dimensions of conflict." (Knight and Johnson 1994: 282) In addition to inducing agreement on the specific dimension of conflict, John Dryzek and Christian List (2013) have argued that a prior stage of deliberation before aggregating individual votes may generate escape routes to the usual impossibility results of social choice theory – in particular, deliberation may induce individuals to form what are known as "single-peaked preferences," which prevents majority cycles that may afflict society in the aggregative stage. (Dryzek and List 2003; see also List et al. 2013)

In sum, there is an abundance of work in deliberative democratic theory that attempts to salvage democratic justification and legitimacy from the impossibility results of social choice theory. Yet, in contrast to social choice theory, there has been relatively little work done to contruct a formal theory of democratic deliberation itself. As Knight and Johnson (1997) observe, unlike the "systematic analysis of the normative and analytical properties of voting procedures" of social choice theory, "[n]o comparable analysis exists for deliberative democracy." (Knight and Johnson 1997: 282) This is unfortunate, because formal analysis has the potential to generate important insights into democratic deliberation. For example, Dimitri Landa and Adam Meirowitz have argued that by largely ignoring the key insights discovered in the game-theoretic analysis on communication, deliberative democratic theorists have been blind to the effect of different deliberative environments on the structure of individual incentives, which in turn may either promote or hinder successful deliberation among strategic actors. (Landa and Meirowitz 2009) The literature is beginning to fill this gap in the theory of deliberative democracy, as there is a growing number scholars who are offering formal theories of democratic deliberation, as well as its cognate notions of democratic justification and legitimacy. (Dietrich and List 2013; Patty 2008; Patty and Penn 2011, 2014; Perote-Peña and Piggens 2015)

This paper presents a formal theory of democratic deliberation that aims to both join and complement this growing line of research. Our contribution is

to explicitly model the process by which reasons are given and arguments are made, and to capture the dynamic nature of deliberation. Our formal theory will also take into consideration the strategic aspects of political debate, in which those who have conflicting interests try to reach common grounds via reasoned argumentation within public deliberation.

2 Overview of Our Theory

Here, we give an overview of our formal theory of deliberation.[4] Our focus is on the dynamics and outcomes of three different modes of deliberation: (i) *myopic discussion*, in which positions on an issue are compared and subject to argument in a relatively free-flowing manner; (ii) *constructive discussion,* in which deliberation follows an argument-climbing dynamic, and (iii) *debate* between interested parties, each of whom seeks to employ rhetorical tactics to her advantage. The analysis of debate layers the structure of a non-cooperative game on top of the framework of deliberation, and in this case, the evolution of the debate does not arise mechanically as the result of behavioral or cognitive assumptions imposed on the participants; instead, it is derived endogenously, from the equilibrium incentives of the participants.

The modeling framework takes as primitive notions: (a) a set of positions to be considered, (b) a set of arguments that can be made for or against different positions, and (c) an assessment of the effectiveness of these arguments. If two people are making a joint decision about which model of car to purchase, for example, then one can imagine many different reasons (*e.g.,* price, fuel economy, safety, performance, reliability, etc.) that could be used within an argument that supports a decision to buy one car over another. Here, we do not consider the specific verbal formulations that different arguments can possibly take, but rather, we will generally conceive an argument as a case of using a particular reason to support one alternative over another. The effectiveness of arguments is modeled by a "set-valued relation" on the set of positions, where given any positions x and y, the set $p(x, y)$ consists of the set of arguments/reasons that are effective for x against y. From fundamental assumptions about the effectiveness of arguments, it is deduced that each argument can be viewed as a ranking of positions. Three different deliberative dynamics are then considered and their properties are examined.

First, we consider what we call a *myopic discussion,* in which positions are considered according to an exogenous protocol and arguments applied to them. An initial status quo position is given, and this evolves in a context-free way: if the position-argument pair given by the protocol are such that the position

[4]Among previous works, that closest in spirit to our analysis is Patty (2008). There, John Patty presents a formal theory of what he calls an "argument-based collective choice," in which an "argument" is defined as a series of links that ultimately connect a (widely accepted) first principle to a given collective action. (Patty 2008: 386) Patty then explores and applies different notions of stability to his theory of arguments.

is superior to the status quo with respect to the argument, then it becomes the new status quo. To illustrate, consider the problem of buying a car, and suppose that among the many relevant criteria, a luxury sedan is superior to a minivan, which in turn is superior to a sports car, with respect to fuel economy. If the initial status quo is, say, a minivan, then this can be replaced by the luxury sedan using the fuel economy argument. At a later point in the discussion, if the status quo is the sports car, then it could be replaced by either the luxury sedan or the minivan by the same fuel economy argument. In a myopic discussion, nothing prevents the sports car from getting replaced by the minivan, which was the initial status quo; for this mode of deliberation, it does not matter that the luxury sedan has already been argued for on the basis of fuel economy and has already replaced the minivan on this account in the previous rounds of discussion. A myopic discussion is thus susceptible to cycles. We show that myopic discussion can be conclusive (*i.e.,* converges on a single position) only under restrictive conditions, and that the long run outcomes of myopic discussion can be highly indeterminate. The result is that a myopic discussion, despite being a form of democratic deliberation, fails to serve many ideals of deliberative democracy.

Next, we provide a model of what we call *constructive discussion,* in which positions are again considered according to an exogenous protocol, but, unlike myopic discussions, once a position x is justified as status quo via a particular argument a, no other position y can be justified via the same argument unless it is superior to x according to that argument. In the car purchase context, assume the initial status quo is the sports car. This can be replaced by the minivan using the fuel economy argument, but if so, we assume that the minivan cannot later be justified again by the fuel economy argument: the idea is that previously in the discussion, the minivan was argued for on the basis of fuel economy, and that was not sufficient to conclude the discussion, so the fuel economy argument can only be used to argue for a better position – in this case, the luxury sedan. This precludes the possibility of cycles that plagued myopic discussion, and it implies that constructive discussions follow an "argument-climbing" dynamic. We show that a constructive discussion must eventually conclude with a position that is top ranked according to some argument, lending the outcome of a constructive discussion a strong justification according to at least one reason or criterion. However, we also show that these outcomes are path dependent: under general conditions, every position that is top ranked according to some argument can be supported as the conclusion of a constructive discussion. The upshot is that although a constructive discussion does better than a myopic discussion (specifically, it concludes with an unanimous agreement on a single position), it still fails to confer full democratic justifcation or legitmacy as the conclusion reached through construtive discussion is essentially arbitrary.

Finally, we present a model of *debate*, in which two participants have diametrically opposed preferences and the protocol is formed endogenously as the equilibrium path of play of a two-player, zero-sum, extensive form game of per-

fect information. We show that there is a unique Nash equilibrium outcome of this game, *a fortiori,* this is also the unique subgame perfect equilibrium outcome. Specifically, assume for simplicity that the number of reasons available to the participants is odd. Then there is a unique position, say x^*, that is top ranked for some argument and such that no more than half of the arguments have top ranked position preferred to x^* by participant 1, and no more than half of the arguments have top ranked positions preferred to x^* by participant 2. Roughly speaking, x^* is middle-ranked by both players among the positions that appear at the top of some argument. We show that this *compromise position* is the unique equilibrium outcome of the debate game and is thus the unique conclusion of any debate, irrespective of the initial status quo. As a mode of democratic deliberation, a debate has many attractive properties; in particular, the outcome of a debate is unique and path independent, has strong justification according to at least one reason or argument, and represents fair and equal concessions on the parts of the participants. Surprisingly, far from resulting in conflict and extreme polarization, it is the addition of diametrically opposed interests as well as the added the strategic incentives among the participants that render a debate to meet the many lofty ideals of deliberative democracy.

3 A Formal Model of Arguments

One important premise of deliberative democratic theory is that it is the force of better reasons and better arguments that determine the legitimacy of political outcomes. "Deliberation is *reasoned,*" says Cohen "in that the parties to it are required to state their reasons for advancing proposals, supporting them, or criticizing them. They give reasons with the expectation that those reasons (and not, for example, their power) will settle the fate of their proposal." (Cohen 1997a: 74) During democratic deliberation, "no force except that of the better argument is exercised." (Habermas 1975: 108) In this section, we model the most basic and important component of a theory of democratic deliberation: arguments (or equivalently, for us, reasons).

Let A be any nonempty, finite set, which we will interpret as a set of possible arguments or reasons, and let X be a set consisting of at least two positions; in general, X may be infinite, but we assume it is finite for many of our results. A reason in our model is simply a standard or a criterion that one may use to construct an argument in support of a given position against another position. Since a reason will only be used in our theory along with an argument in support of a given position against another, we will use the terms "reason" and "argument" interchangeably, depending on the context.

A *binary relation* on X is any subset $P \subseteq X \times X$ of ordered pairs of elements from X; as is customary, we write xPy instead of $(x, y) \in P$. A binary relation P on X is *asymmetric* if for all $x, y \in X$, not both xPy and yPx; it is *transitive* if for all $x, y, z \in X$, xPy and yPz together imply xPz;

and it is *total* if for all distinct $x, y \in X$, either xPy or yPx. As is customary, we write $xPyPz$ to denote the conjunction xPy and yPz. A *partial order* is a binary relation that is asymmetric and transitive, and a *linear order* is a partial order that is total. A position x is *maximal* with respect to P if there is no position y such that yPx. It is well-known that if X is finite and P is a partial order, then there is at least one maximal element of P.

A *set-valued relation* on X is a mapping $p\colon X \times X \to 2^A$ that associates a set $p(x, y) \subseteq A$ of reasons/arguments to each ordered pair $(x, y) \in X \times X$ of positions. We can view a binary relation as a mapping $P\colon X \times X \to 2^{\{1\}}$, where xPy holds if and only if $P(x, y) = \{1\}$, and in this way, set-valued relations generalize the usual concept of a binary relation. In the present context, we interpret $p(x, y)$ as the set of reasons that can be effectively used to support the position x over position y. We maintain the assumption that p is *asymmetric*, in the sense that for all $x, y \in X$, we have $p(x, y) \cap p(y, x) = \emptyset$; in words, no argument/reason that is effective for x against y is effective for y against x. In particular, for all $x \in X$, we have $p(x, x) = \emptyset$. That is, no argument/reason can be used simultaneously to both support and reject a position against itself. In this sense, our set-valued relation p is *irreflexive.*

A set-valued relation p is *total* if for all distinct positions $x, y \in X$ and all arguments $a \in A$, either $a \in p(x, y)$ or $a \in p(y, x)$, *i.e.,* for all distinct $x, y \in X$, $p(x, y) \cup p(y, x) = A$. In other words, if a set-valued relation p is total, then this means that there are no redundant or superfluous arguments/reasons in A: given any two distinct positions $x, y \in X$, any reason $a \in A$ is either an argument for x against y or an argument for y against x. We say p is *transitive* if for all $x, y, z \in X$, we have $p(x, y) \cap p(y, z) \subseteq p(x, z)$. In other words, transitivity of p means that whenever a given reason $a \in A$ is an effective argument for x against y and is also an effective argument for y against z, then the same reason a is an effective argument for x against z as well.

Next, having introduced the concept of set-valued relation, we establish that the properties of p can be analyzed by means of a collection of binary relations on the set of positions. For each argument $a \in A$, define the relation P^a on X as follows: xP^ay if and only if $a \in p(x, y)$. Clearly, as p is asymmetric, each relation P^a is asymmetric. Our first proposition shows that other properties of p are mirrored in these relations as well: p is total if and only if each P^a is total, and p is transitive if and only if each P^a is transitive, *i.e.,* a partial order.

Theorem 1: The set-valued relation p is total if and only if for all $a \in A$, P^a is total. Moreover, p is transitive if and only if for all $a \in A$, P^a is transitive.

Proof: First, assume p is total, and consider any distinct $x, y \in X$. Since p is total, we have $a \in p(x, y)$ or $a \in p(y, x)$, which implies xP^ay or yP^ax, and thus P^a is total. Conversely, assume each P^a is total, and consider distinct $x, y \in X$ and $a \in A$. Since P^a is total, either xP^ay or yP^ax, which implies $a \in p(x, y)$ or $a \in p(y, x)$, and thus p is total. Second, assume p is transitive, and consider any $x, y, z \in X$ and $a \in A$ such that xP^ayP^az. Then $a \in p(x, y) \cap p(y, z)$.

Since p is transitive, we have $a \in p(x, z)$, which implies $xP^a z$, and thus P^a is transitive. Conversely, assume each P^a is transitive, and consider $x, y, z \in X$ and $a \in p(x, y) \cap p(y, z)$. Then $xP^a yP^a z$, and transitivity of P^a implies $xP^a z$, *i.e.,* $a \in p(x, z)$. We conclude that $p(x, y) \cap p(y, z) \subseteq p(x, z)$, and thus p is transitive. □

Recall that an asymmetric relation is a linear order if it is total and transitive. Combining the observations of Theorem 1, we conclude that p is total and transitive if and only if each P^a is a linear order. That is, whenever p is total and transitive, each reason $a \in A$ can be seen as a standard or criterion according to which we are able to totally order all the different positions in X. Next, we illustrate this with an example:

Example (Which Car Should We Buy?): There are three alternatives under consideration for the purchase of a new car – a luxury sedan (L), a minivan (V), and a sports car (S) – and three criteria are relevant – fuel economy (f), cost (c), and performance (p). Then the set of positions and set of arguments are

- $X = \{L, V, S\}$
- $A = \{f, c, p\}$.

Assuming that each argument is total (so that no two types of car are equal by any criterion) and transitive, Theorem 1 implies that we can summarize the effectiveness of the arguments by three linear orders, P^f, P^c, and P^p. For example, we might suppose the following rankings of cars by the three criteria:

P^f	P^c	P^p
L	V	S
V	S	L
S	L	V

Thus, the luxury sedan is the most fuel efficient (followed by the minivan and the sports car), the minivan is the cheapest (followed by the sports car and the luxury sedan), and the sports car has the best performance (followed by the luxury sedan and the minivan). □

Let us say that a position x is *unassailable* if there is no position that is superior to it on any argument, *i.e.,* for all $y \in X$, $p(y, x) = \emptyset$, and we denote the set of unassailable positions by UA. An unassailable position, if any, would be a very strong candidate for a collective agreement, as we would not be able to find a different position that is superior to it on *any* grounds. However, such a position may not be available – in our previous "Which Car Should We Buy?" example, there is no unassailable position – and the need for deliberation may be most pressing precisely when such a compelling position cannot be found, *i.e.,* when UA is empty.

Given two positions, x and y, we say x *dominates* y, and write $x\overline{P}y$, if both of the following hold: $p(x,y) \neq \emptyset$, and for all $z \in X$, we have $p(z,x) \subseteq p(z,y)$. That is, there is an argument for x over y, and for every position z, every argument for z over x is also an argument for z over y. If x dominates y, then this implies that there exists no reason or argument according to which y is better than x. In this case, it is clear that y would not be a plausible choice: in order for y to be chosen, some argument would have to eliminate x, but then it would eliminate y as well. We say position x is *undominated* if there is no y that dominates it, and we let UD denote the set of undominated positions, *i.e*, $UD = \{x \in X | \nexists y \in X \text{ such that } y\overline{P}x\}$. Note that if a position is unassailable, then it is undominated. This is because in order for some position to dominate another position there has to be an argument according to which the former is superior to the latter, but, by the definition of an unassailable position, there exists no such argument. Hence, $UA \subseteq UD$.

Next, we note that the dominance relation $\overline{P}$ is a partial order, implying that when X is finite, an undominated position exists, and we give a characterization of the undominated positions: a position x is undominated if for every distinct position y, $p(x,y) \neq \emptyset$, and assuming p is total, the converse holds as well.

Theorem 2: The dominance relation $\overline{P}$ is a partial order; and thus if X is finite, then there is an undominated position. Moreover, given any $x \in X$, if $p(x,y) \neq \emptyset$ for every $y \in X \setminus \{x\}$, then x is undominated. Assuming p is total, then for all $x, y \in X$, $x\overline{P}y$ holds if and only if for all $a \in A$, $xP^a y$; and thus if x is undominated, then for every $y \in X \setminus \{x\}$, we have $p(x,y) \neq \emptyset$.

Proof: Note that $\overline{P}$ is asymmetric, for $x\overline{P}y$ and $y\overline{P}x$ would imply the existence of $a \in p(x,y) \subseteq p(x,x)$ (setting $z = x$ in the definition of $y\overline{P}x$ for the inclusion), contradicting asymmetry of p. For transitivity, consider any positions $x, y, z \in X$ and assume $x\overline{P}y\overline{P}z$. Then there exists $a \in p(x,y) \subseteq p(x,z)$ (now using $y\overline{P}z$ for the inclusion), so that $p(x,z) \neq \emptyset$. Consider any position s and argument $a \in p(s,x) \subseteq p(s,y) \subseteq p(s,z)$. Then $a \in p(s,z)$, and we conclude that $p(s,x) \subseteq p(s,z)$, and that $x\overline{P}z$. Thus, $\overline{P}$ is a partial order. If X is finite, it follows immediately that $\overline{P}$ admits a maximal element, *i.e.,* an undominated position. Now, consider any position $x \in X$, and assume that for all $y \in X \setminus \{x\}$, we have $p(x,y) \neq \emptyset$, and consider any $y \in X$. If y dominated x, then setting $z = x$, we would have $\emptyset \neq p(x,y) \subseteq p(x,x) = \emptyset$, a contradiction; thus, y does not dominate x, and we conclude that x is undominated. Last, assume p is total, and suppose that a position x is undominated but for some position y, we have $p(x,y) = \emptyset$. Since p is total, this implies $p(y,x) = A$. Clearly, $p(y,x) \neq \emptyset$. Now consider any position z. Since $p(z,y) = p(z,y) \cap p(y,x) \subseteq p(z,x)$, by transitivity of p, it follows that y dominates x, a contradiction. □

Assuming p is total, Theorem 2 implies that if a position x is undominated, then for every other feasible position $y \in X \setminus \{x\}$, there will be at least one

effective reason $a \in A$ that we can use to construct an argument in support of x against y. In our car example, one can easily check that none of the car types are dominated. Accordingly, for any given car type in X, we can always find a reason/argument that we can invoke to defeat some other car type in X. For example, to support the prospect of buying a minivan, one could argue that the minivan has better fuel economy than the luxury sedan, and it costs less than the sports car. Similarly, in support of the sports car, one could argue that it costs less than the luxury sedan and drives better than the minivan. Lastly, in support of the luxury sedan, one could argue that it drives better than the minivan and has better fuel economy than the sports car.

Using Theorem 1, when X is finite, each linear order has a position x^a uniquely ranked at the top of the ordering. Such a position is clearly undominated, and these positions will possess a more stringent stability property than that described in Theorem 2: for all $a \in A$ and all $y \in X \setminus \{x^a\}$, we have $a \in p(x^a, y)$. In terms of our car example, the luxury sedan is best in terms of fuel economy, the sports car is best in terms of performance, and the minivan is best in terms of cost. Hence, fuel economy is an effective argument to support the luxury sedan against both the sports car and the minivan; performance is an effective argument to support the sports car against both the luxury sedan and the minivan; and cost is an effective argument to support the minvan against both the luxury sedan and the sports car.

Now, let us define the binary relation P^* on X such that for all positions $x, y \in X$, xP^*y holds if and only if $p(x, y) \neq \emptyset$, *i.e.,* there is at least one argument in favor of x over y. Define the transitive closure of P^*, denoted P^∞, as follows: $xP^\infty y$ if and only if there exist a natural number k and positions $x_1, \ldots, x_k \in X$ such that $xP^*x_1P^* \cdots x_{k-1}P^*x_k = y$. Note that the transitive closure is transitive, but not necessarily asymmetric. We say x is *maximal* with respect to P^∞ if and only if for all $y \in X$, $yP^\infty x$ implies $xP^\infty y$; and we define the *top cycle,* denoted TC, as the set of maximal elements of P^∞.

One may think of the top cycle as a "first cut" of candidate positions to which we should restrict our choice, when forced to take a position. Positions in the top cycle have a *minimal degree* of plausibility, in the sense that they are the positions that can be eventually and repeatedly reached after suitable application of arguments. Because the definition of the top cycle is permissive, it will likely contain any compelling position; for example, every unassailable position belongs to the top cycle, *i.e.,* $UA \subseteq TC$. The top cycle has the desirable property that it is generally non-empty even when there are cycles. However, the problem of the top cycle is that it can be very large resulting in indeterminacy or the lack of sharp prescription, which may be problematic in many contexts in which political deliberation takes place.

Next, we provide a decomposition of the top cycle into separate components. Formally, given a subset $Y \subseteq X$ of positions, we write $YP^\infty x$ if every element of the set bears the transitive closure, P^∞, to x, *i.e.,* for all $y \in Y$, $yP^\infty x$. Then a nonempty set Y of positions is a *component* if both of the following conditions hold: for all $x \in Y$, we have $(Y \setminus \{x\})P^\infty x$; and there is

no superset of $Z \supsetneq Y$ such that for all $y \in Y$, $ZP^{\infty}y$. Roughly, a component is maximal (with respect to set inclusion) among sets that bear P^{∞} (*i.e.,* the transitive closure of P^{*}) to each of their elements. We show that the top cycle consists of the union of components; moreover, if X is finite and a position $x \in X \setminus TC$ does not belong to the top cycle, then there is a component Y such that $YP^{\infty}x$. Note that the result implies that when X is finite, the top cycle is nonempty.

Theorem 3: The top cycle is the union of components, *i.e.,* $TC = \bigcup\{Y \mid Y \text{ is a component}\}$. Moreover, if X is finite, then for every $x \in X \setminus TC$, there is a component Y such that $YP^{\infty}x$. Finally, if p is total, then TC consists of a single component, and $UA \subseteq UD \subseteq TC$.

Proof: First, let $x \in TC$ be a position in the top cycle, and define $Y = \{y \in X \mid yP^{\infty}x\} \cup \{x\}$. If $Y = \{x\}$, then it is clearly a component, so assume Y contains at least one position distinct from x. Note that for all $y \in Y \setminus \{x\}$, because x is maximal with respect to P^{∞}, we have $xP^{\infty}y$. Now consider $y \in Y$ and $z \in Y \setminus \{y\}$. If $z = x$, then we have shown $zP^{\infty}y$; and otherwise, we have $zP^{\infty}xP^{\infty}y$, which again implies $zP^{\infty}y$. Since $z \in Y$ is arbitrary, we thus have $(Y \setminus \{y\})P^{\infty}y$. Given $y \in Y$ and $z \in X \setminus Y$, we cannot have $zP^{\infty}y$, for that would imply $zP^{\infty}x$, which is impossible since $z \notin Y$. We conclude that Y is a component. Thus, the top cycle is contained in the union of components. Second, let Y be any component, let $x \in Y$ be a position in Y, and consider any $y \in X$ such that $yP^{\infty}x$. If $y = x$, then $xP^{\infty}y$ follows immediately, so assume $y \neq x$. If $y \in Y$, then by definition of a component, we have $(Y \setminus \{y\})P^{\infty}y$, which implies $xP^{\infty}y$. To show that x belongs to the top cycle, it then suffices to rule out $y \in X \setminus Y$. Suppose toward a contradiction that $y \notin Y$. In case $Y = \{x\}$, we have $(Y \cup \{y\})P^{\infty}x$, contradicting the assumption that Y is a component, and, hence, the largest set that bears P^{∞} to each of its elements except itself. In the remaining case that Y contains positions distinct from x, given any $z \in Y$, we have $yP^{\infty}xP^{\infty}z$, which implies $yP^{\infty}z$. Again, we arrive at $(Y \cup \{y\})P^{\infty}z$, and since $z \in Y$ is arbitrary, this contradicts the assumption that Y is a component. We conclude that x belongs to the top cycle, and, therefore, the top cycle consists of the union of components.

Next, assume X is finite, and consider $x \in X \setminus TC$. Let $Y = \{y \in X \mid yP^{\infty}x\}$, which is nonempty and finite. We claim that there is a maximal element of P^{∞} in Y, *i.e.,* a position $y^{*} \in Y$ such that for all $z \in Y$, if $zP^{\infty}y^{*}$, then $y^{*}P^{\infty}z$. Indeed, we can define $\tilde{P}$ as the asymmetric part of P^{∞}, so that for all $s, t \in X$, $s\tilde{P}t$ if and only if $sP^{\infty}t$ but not $tP^{\infty}s$. It is then straightforward to check that $\tilde{P}$ is a partial order, and thus it admits maximal elements in Y, and these fulfill the claim. Define $Z = \{z \in X \mid zP^{\infty}y^{*}\} \cup \{y^{*}\}$. By the initial argument of the proof, it follows that Z is a component, and by transitivity of P^{∞}, we have $ZP^{\infty}x$, as required.

Last, assuming p is total, consider any components Y and Z. If these components are singleton and consist of the same position, then clearly $Y = Z$. Otherwise, we can choose distinct positions $y \in Y$ and $z \in Z$. Since p is total,

we can assume without loss of generality that yP^*z. Then for all $x \in Z$, we have $(Y \cup Z)P^\infty zP^\infty x$, which implies $(Y \cup Z)P^\infty x$. By definition of a component, it follows that $Y \cup Z \subseteq Z$, *i.e.,* $Y \subseteq Z$. Then $zP^\infty y$, and a symmetric argument implies $Z \subseteq Y$, *i.e.,* $Y = Z$. We conclude that the top cycle consists of a single component. Now, consider any undominated position $x \in UD$. By Theorem 2, it follows that for all $y \in X \setminus \{x\}$, we have xP^*y, and thus $xP^\infty y$, and this implies $x \in TC$. We conclude that $UD \subseteq TC$. □

In the next sections, we employ this apparatus of reasons/arguments to provide a formal analysis of the deliberative dynamics of three different modes of democratic deliberation: myopic discussion, constructive discussion, and (strategic) debate.

4 Myopic Discussion

A frequently asked question among scholars of deliberative democratic theory is whether the participants in deliberation will eventually reach rational consensus or unanimous agreement at the end of deliberation. Many deliberative democrats have suggested that deliberative democracy should, at least at the theoretical level, ideally target unanimous agreement as its ultimate aim. For instance, Cohen explains that "ideal deliberation aims to arrive at a rationally motivated consensus." (Cohen 1997a; 75) Elster explains that "not optimal compromise, but unanimous agreement is the goal of politics on this view." (Elster 1997: 12; see also Habermas 1990) Gaus calls this requirement "the Regulative Ideal of Real Political Consensus" and explains that the normative commitments of many deliberative democrats have led them to endorse such a view. (Gaus 1997: 206)

However, many critics have argued that full consensus or unanimous agreement is unlikely to be achieved in realistic circumstances, especially, in modern pluralistic societies. According to Knight and Johnson, "any imaginable human population is diverse across multiple, overlapping dimensions including material interests, moral and ethical commitments, and cultural attachments," implying that "disagreement and conflict are unavoidable." (Knight and Johnson 2007: 47) Furthermore, even among reasonable people, having more reasoned deliberation may not resolve such a disagreement due to what Rawls calls "the burdens of judgment" (Rawls 1999/2003: 54-58) This means that "even assuming unlimited time for discussion, unanimous and rational agreement might not necessarily ensue." (Elster 1997: 14) Some critics have even gone further to argue that deliberation "only rarely eliminates differences of opinion on matters of politics" and that it may actually produce even "more disagreement and diversity of opinion." (Christiano 1997: 264)

In many political circumstances, time is limited, and a collective decision may have to be made after some duration of deliberation, even if that deliberation fails to reach unanimous agreement. Many deliberative democrats tend to

advert back to aggregative mechanisms when this happens. For instance, even Cohen, who deems unanimous agreement as the ultimate aim of ideal deliberation, acknowledges that "[e]ven under ideal conditions there is no promise that consensual reasons will be forthcoming" and "[i]f they are not, then deliberation concludes with voting, subject to some form of majority rule." (Cohen 1997a: 75) Here, the reliance on aggregative mechanisms is treated by deliberative democrats not as something desirable in itself, but as a "necessary evil" that must be resorted to for practical purposes.

In this section, we introduce the concept of discussion and explore the dynamics of one particular type of discussion, namely, myopic discussion. We then examine whether myopic discussions can fulfill one of the ideals of deliberative democracy by leading the participants to reach a unanimous agreement on a single position. According to Gutman and Thomson (2004), one of the major characteristics of a deliberative democracy is that its process is dynamic; that is, "[a]lthough a decision must stand for some period of time, it is provisional in the sense that it must be open to challenge at some point in the future." (Gutman and Thomson 2004: 9) To incorporate this feature, our model of discussion will consist of a sequence of rounds in which positions are retained or replaced on the basis of arguments exchanged during each round of the discourse.

We imagine discussion proceeding sequentially over an unbounded number of rounds such that in each round m, there is a current status quo position z^m, which is subject to a challenge by position x^m and argument a^m. Formally, a sequence $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ in $X \times A \times X$ is a *discussion* if for all m, we have:

- $z^{m+1} \in \{x^m, z^m\}$,
- $z^{m+1} = z^m \neq x^m$ implies $a^m \notin p(x^m, z^m)$.

If $x^m = z^{m+1}$, then we say z^m is *justified* by argument a^m, and if $x^m = z^{m+1} \neq z^m$, then x^m is *inserted* by the argument a^m; the difference is that if position x^m is the status quo in round m, *i.e.,* $z^m = x^m$, then it can be justified by an argument as the new status quo in round $m+1$, but not technically inserted (as it was the status quo previously). The sequence $\mathfrak{P} = \{(x^m, a^m)\}_{m=1}^{\infty}$ of position-argument pairs introduced to challenge the status quo position is a *protocol,* and we say the discussion is *open* if the following holds: for all position-argument pairs (x, a) and all rounds m, if there is a position y such that $xP^a y$, then there exists $n \geq m$ such that $(x^n, a^n) = (x, a)$. A position-argument pair (x, a) for which there exists a y with $xP^a y$ is said to be *potentially effective,* so that an open protocol is one in which each potentially effective position-argument pair appears infinitely often. According to Cohen, a defining characteristic of a deliberative democracy is "continuity," meaning that when there are no time constraints, the process of deliberation should ideally continue "into the indefinite future." (Cohen 1997a: 72) Our assumption that discussions are infinite is meant to reflect this open-ended nature

of this type of discourse, and our assumption that each position-argument pair appears infinitely often in an open discussion implies that any position is vulnerable to replacement by any argument for which it is not top-ranked.

Thus, a discussion must follow a given protocol, and a new position can replace the status quo only if it is superior according to the currently salient argument. Of course, the second condition in our definition of discussion gives only a necessary condition for replacement of the status quo, so the definition is too broad: given any protocol, we could select an arbitrary position z, even one ranked last according to all arguments, and set $z^m = z$ for all m. What is needed is a restriction on discussion that includes a sufficient condition for replacement of the status quo, ruling out discussions that stabilize at implausible positions.

Our first step in this direction is to define a *myopic discussion* as an open discussion such that if position x^m is superior to z^m according to argument a^m, then it becomes the status quo in the next round. Formally, it is an open discussion $\{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ such that for all m, we have:

$$z^{m+1} = \begin{cases} x^m & \text{if } a^m \in p(x^m, z^m), \\ z^m & \text{else.} \end{cases}$$

In such a discussion, a position may replace the status quo by any argument for which it is superior, regardless of the history of discussion. It may be, for example, that in some round m, position x is inserted by an argument a against status quo z, and in some later round, the same position is inserted by the same argument against the same status quo.

Because they evolve in a context-free way, myopic discussions permit cycles and can, in principle, repeat *ad infinitum.* To illustrate this, we return briefly to the car example, and we provide a myopic discussion that cycles through the various models of car. Clearly, such a discussion is not constructive, a point to which we return in short order.

Example (Myopic Discussion May Generate Cycles): In the car purchase example, set the initial status quo equal to the sports car, *i.e.,* $z^1 = S$, and consider the following discussion,

x^m	V	L	S	V	L	S	$\cdots$
a^m	f	f	c	c	p	p	$\cdots$
z^{m+1}	V	L	S	V	L	S	$\cdots$

and so on, repeating thereafter with periodicity six. Here, the length of the cycle reflects the fact that there are six potentially effective position-argument pairs, each of which must appear infinitely often in the protocol. Indeed, the minivan is ranked above the sports car in terms of fuel economy, and the luxury sedan is ranked above the minivan on that basis, and these position-argument pairs appear at the beginning of the protocol. We then use the pairs (S, c) and (V, c), and the last to rounds in the above segment use the pairs (L, p) and

(S, p). Thus, each position-argument pair (x, a) such that x is not bottom-ranked according to a appears infinitely often in the protocol, as required for an open discussion. In each round m, the status quo faces a position x^m that is superior according to a^m, and the new position is inserted as status quo, fulfilling the definition of myopic discussion. □

We can summarize the long run dynamics of a discussion $\mathfrak{D}$ by the *limit set,* denoted $\Lambda(\mathfrak{D})$, of positions that appear as status quo infinitely often in the discussion; formally, we specify that $x \in \Lambda(\mathfrak{D})$ if and only if for all n, there exists $m \geq n$ such that $x = z^m$. Assuming the set of positions is finite, since all positions outside the limit set appear in the discussion only finitely many times, a discussion eventually reaches a period after which the only positions inserted as status quo are positions in the limit set. Since a discussion is a sequence (*i.e.,* there is an infinite number of rounds), every discussion will have a non-empty limit set: $\Delta(\mathfrak{D}) \neq \emptyset$. However, the existence of a non-empty limit set does *not* imply that our myopic discussion will reach unanimous agreement. If the limit set contains more than one position, then a myopic discussion cycles through the different positions in the limit set, representing perpetual disagreement in the myopic discussion.

The next result characterizes the long run outcomes of a myopic discussion. In general, not much can be said, but for every position x outside the limit set and every argument a, there must be some position in the limit set that is not vulnerable to x by argument a; Corollary 1 will extract a further implication of this observation for the conclusiveness of myopic discussions. When p is total, however, the implications are much sharper: every myopic discussion eventually reaches the top cycle and remains thereafter.

Theorem 4 (Long Run Outcomes of Myopic Discussion): Assume X is finite. For every myopic discussion $\mathfrak{D}$ and every position $x \in X \setminus \Lambda(\mathfrak{D})$ and every argument $a \in A$, there exists $y \in \Lambda(\mathfrak{D})$ such that $a \notin p(x, y)$. Moreover, if p is total, then for every myopic discussion $\mathfrak{D}$, we have $\Lambda(\mathfrak{D}) \subseteq TC$.

Proof: First, assume X is finite, and let $\mathfrak{D}$ be a myopic discussion. Consider any position $x \in X \setminus \Lambda(\mathfrak{D})$ and argument a, and suppose toward a contradiction that for all $y \in \Lambda(\mathfrak{D})$, we have $a \in p(x, y)$. For each $z \in X \setminus \Lambda(\mathfrak{D})$, there exists m_z such that for all $n \geq m_z$, we have $z^n \neq z$. Since X is finite, we can let $m = \max\{m_z \mid z \in Z \setminus \Lambda(\mathfrak{D})\}$, which means that for all $n \geq m$, we have $z^n \in \Lambda(\mathfrak{D})$. Since $xP^a z^m$, the definition of an open discussion implies that (x, a) appears infinitely often in $\mathfrak{D}$, *i.e.,* there exist arbitrarily large $n \geq m$ such that $(x^n, a^n) = (x, a)$. For each such n, we have $a \in p(x, z^n)$ by supposition, and thus the status quo in round $n + 1$ is $z^{n+1} = x$. But this implies that $x \in \Lambda(\mathfrak{D})$, a contradiction. Thus, the first part of the theorem holds. Second, assume p is total, and consider any myopic discussion $\mathfrak{D}$. We claim that for all $x \in TC$ and all $y \in X \setminus TC$, we have $p(x, y) = A$. Indeed, since x is in the top cycle and y is not, we cannot have yP^*x, and thus $p(y, x) = \emptyset$. Since p is total, we have $p(x, y) = A$, as claimed. Now, choose any $x \in TC$, and note

that since p is total, there is a position y such that xP^*y, and by definition of open discussion, there exists m such that $x^m = x$. By the preceding claim, it follows that $z^{m+1} \in TC$. Indeed, if $z^m \in TC$, then this holds trivially; and if $z^m \notin TC$, then the claim implies $a^m \in p(x, z^m)$, so that $z^m = x \in TC$. Then for all $n > m$, the status quo remains in the top cycle, which implies $\Lambda(\mathfrak{D}) \subseteq TC$. □

The implications of Theorem 4 are limited for general p, and it uses the assumption that p is total to conclude that the long run outcomes of myopic discussion are contained in the top cycle. The next example demonstrates that this ancillary assumption is needed for the result: without it, a myopic discussion may cycle through a limit set $\Lambda(\mathfrak{D})$ that does not intersect the top cycle.

Example (Myopic Discussion May Not Reach Top Cycle): Assume there are four positions, $X = \{A, B, C, D\}$, and four arguments, $A = \{a, b, c, d\}$. For this example, we assume a simple structure such that A is superior to B by argument a, B is superior to C by argument b, C is superior to D by argument c, and D is also superior to A, but by argument d instead. That is,

$$\begin{aligned} p(A, B) &= \{a\} & \qquad p(C, A) &= \{c\} \\ p(B, C) &= \{b\} & \qquad p(D, A) &= \{d\}, \end{aligned}$$

and $p(x, y) = \emptyset$ for all other pairs of positions. Then the relation P^* consists of four ordered pairs, $P^* = \{(A, B), (B, C), (C, A), (D, A)\}$. Clearly, position D is unassailable, and we have $DP^*AP^*BP^*C$; in particular, for every other position $x \in X \setminus \{D\}$, we have $DP^\infty x$ but not $xP^\infty D$. Thus, the top cycle is the singleton $TC = \{D\}$. We construct a discussion $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^\infty$ with initial status quo $z^1 = A$ and protocol as follows:

x^m	C	D	B	A	$\cdots$
a^m	c	d	b	a	$\cdots$
z^{m+1}	C	C	B	A	$\cdots$

and so on, repeating with periodicity equal to four thereafter. By definition of myopic discussion, the status quo evolves as follows: $z^2 = C$ is inserted via argument $a^1 = c$, $z^3 = C$ remains because $a^2 = c \notin p(D, C)$, $z^4 = B$ is inserted via argument $a^3 = b$, $z^5 = A$ is inserted via argument $a^4 = a$, and so on. Thus, the limit set of the discussion is $\Lambda(\mathfrak{D}) = \{A, B, C\}$, which does not intersect the top cycle. □

Recall that our goal is to examine whether the dynamics of myopic discussion can satisfy deliberative democracy's ideal of unanimous agreement. Although we model a discussion as an infinite sequence $\{(x^m, a^m, z^m)\}_{m=1}^\infty$, this is not to say that a discussion cannot be resolved in a finite amount of time. We say a discussion is *conclusive* if there is some round after which the status quo is never revised, *i.e.,* there exists m such that for all $n \geq m$, we have $z^n = z^m$, in which case it *concludes* with the position z^m. For a conclusive

discussion, the continuation for an infinite number of rounds is merely a technicality, for the conversation is essentially over once the status quo has reached its final position. When a discussion is conclusive, we can at least say that it meets one desideratum of deliberative democracy: the ideal of unanimous agreement.

Unfortunately, myopic discussion can be conclusive only under stringent conditions, for such a discussion can be conclusive only if it concludes with an unassailable position – and as we have seen with our car example, the set UA may very well be empty. This negative result is an immediate corollary of Theorem 4: if a myopic discussion $\mathfrak{D}$ is conclusive, then $\Lambda(\mathfrak{D})$ consists of a single position, say y, and then the first part of Theorem 4 implies that for every other position x and every argument a, we must have $a \notin p(x, y)$, implying that y is unassailable.

Corollary: Assume X is finite. If a myopic discussion $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ is conclusive, then there exist an unassailable position x and natural number m such that for all $n \geq m$, we have $z^m = x$.

We have shown that a myopic discussion can cycle through the limit set $\Lambda(\mathfrak{D})$ endlessly, and it can be conclusive only under special circumstances. But how bad can it get? When p is total, Theorem 4 implies that these long run outcomes must belong to the top cycle, giving an upper bound on the indeterminacy of myopic discussion, but the next result shows that this bound can be attained: we specify a protocol that reaches the top cycle from any initial status quo and then cycles through the entire top cycle. In fact, we use a protocol that is *rotating,* in the sense that there exist n positions $x_1, \ldots, x_n$ and arguments $a_1, \ldots, a_n$ such that for all m, $(x^m, a^m) = (x_{m \pmod n}, a_{m \pmod n})$. That is, the first n position-argument pairs are $(x^1, a^1), \ldots, (x^n, a^n)$, and the protocol repeats this pattern thereafter.

Theorem 5 (Indeterminacy of Myopic Discussion): Assume X is finite and p is total. There is a rotating protocol $\mathfrak{P} = \{(x^m, a^m)\}_{m=1}^{\infty}$ such that for every myopic discussion $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ for this protocol, the limit set coincides with the top cycle: $\Lambda(\mathfrak{D}) = TC$.

Proof: Assume X is finite and p is total. If the top cycle consists of just one alternative, say x, then since p is total, we have $p(x, y) = A$ for all $y \in X \setminus \{x\}$, and any protocol in which x appears gives us the result. Henceforth, we assume $|TC| \geq 2$. We say a subset $Y \subseteq X$ of positions is Hamiltonian if $Y \subseteq TC$ and there exist distinct positions $x_1, \ldots, x_n \in Y$ such that $Y = \{x_1, \ldots, x_n\}$ and $x_1 P^* x_2 P^* \cdots x_n P^* x_1$. That is, Y is a subset of the top cycle, and the elements of Y can be indexed $x_1, \ldots, x_n$ in such a way that there is a P^*-cycle in line with this indexing; obviously, in this case, we can also re-index positions so that $x_n P^* x_{n-1} P^* \cdots x_1 P^* x_n$. Let $\mathscr{H}$ denote the collection of Hamiltonian sets. By Theorem 3, the top cycle consists of a single component, and since it contains at least two distinct alternatives, say $x, y \in TC$, we then have $(TC \setminus \{y\}) P^\infty y$ and $(TC \setminus \{x\}) P^\infty x$, which yields $x P^\infty y P^\infty x$. This gives us

a P^*-cycle containing both x and y, *i.e.,* alternatives $x_1, \ldots, x_k \in X$ such that $x_1 P^* x_2 P^* \cdots x_k P^* x_1$ with $x = x_i$ and $y = x_j$ for some $i, j = 1, \ldots, k$. By choosing the shortest such path, we ensure that the alternatives $x_1, \ldots, x_k$ are distinct, and then $Y = \{x_1, \ldots, x_k\}$ is Hamiltonian, and thus $\mathscr{H}$ is nonempty. Since it is finite, we can choose a Hamiltonian set $Y \in \mathscr{H}$ that is maximal among $\mathscr{H}$ with respect to set inclusion. We claim that $Y = TC$, for suppose there exists $x \in TC \setminus Y$. We discern two cases.

Case 1: for all $x \in TC \setminus Y$, if xP^*x_i for some $i = 1, \ldots, k$, then xP^*x_j for all $j = 1, \ldots, k$. That is, if any position $x \in TC \setminus Y$ bears P^* to at least one element of Y, then it bears the relation to all elements of Y. For each $y \in Y$, we have $(TC \setminus \{y\})P^\infty y$, so there is some position $x \in TC \setminus Y$ such that xP^*y, and thus for all $i = 1, \ldots, k$, we have xP^*x_i. Case 1.1: there is some $i = 1, \ldots, k$ such that $x_i P^* x$. Without loss of generality, assume $i = 1$, so $x_1 P^* x$. By assumption, we have xP^*x_2, and thus we have $x_1 P^* x P^* x_2 \cdots P^* x_k P^* x_1$, but then $Y \cup \{x\}$ is Hamiltonian, contradicting maximality of Y. Case 1.2: there is no $i = 1, \ldots, k$ such that $x_i P^* x$. Since $(TC \setminus \{x\})P^\infty x$, there exist $y_1, \ldots, y_\ell \in X$ with $y_1 \in Y$ and $y_1 P^* y_2 P^* \cdots y_\ell P^* x$. Since x belongs to the top cycle, all of the alternatives $y_1, \ldots, y_\ell$ belong to the top cycle as well. Let j be the highest index such that $y_j \in Y$, and note that by the assumption of Case 1.2., we have $j < \ell$. Since $y_j \in Y$, we may write $y_j = x_i$ for some $i = 1, \ldots, k$, and thus we have

$$x_1 P^* x_2 P^* \cdots x_i P^* y_{j+1} P^* \cdots y_\ell P^* x P^* x_1,$$

but then $Y \cup \{y_{j+1}, \ldots, y_\ell, x\}$ is Hamiltonian, contradicting maximality of Y.

Case 2: there exist $x \in TC \setminus Y$ and $i, j = 1, \ldots, k$ such that xP^*x_i and not xP^*x_j. Case 2.1: $i > j$. Then without loss of generality, we can choose i to be the lowest index subject to $i > j$ and xP^*x_i. Then it is not the case that xP^*x_{i-1}, and since p is total, this implies $p(x_{i-1}, x) = A$, and in particular, $x_{i-1} P^* x P^* x_i$, and thus we have

$$x_1 P^* x_2 P^* \cdots x_{i-1} P^* x P^* x_i P^* x_{i+1} \cdots x_k P^* x_1,$$

but then $Y \cup \{x\}$ is Hamiltonian, contradicting maximality of Y. Case 2.2: $i < j$. Then we can choose (i, j) to maximize $j - i$ subject to the constraints that $i < j$, xP^*x_i, and not xP^*x_j. Since p is total, this implies $x_j P^* x$, and identifying x_{k+1} with x_1, we can then write $x_j P^* x P^* x_{j+1}$. Thus, again $Y \cup \{x\}$ is Hamiltonian, contradicting maximality of Y. We conclude that $Y = TC$.

Thus, letting the number of elements of the top cycle be n_1, we can index the top cycle set as $TC = \{x_1, \ldots, x_{n_1}\}$ so that $x_{n_1} P^* x_{n_1 - 1} P^* \cdots x_2 P^* x_1 P^* x_{n_1}$. We define a rotating protocol $\mathfrak{P}$ that consists of three phases.

Phase 1: Note that for all $i = 1, \ldots, n_1$, $x_{i+1} P^* x_i$ implies there exists $a_{i+1} \in A$ such that $a_{i+1} \in p(x_{i+1}, x_i)$, where we identify x_{n_1+1} with x_1, so that $a_1 \in p(x_1, x_{n_1})$. For the first rounds $m = 1, \ldots, n_1$ of the protocol, we set $x^m = x_m$ and $a^m = a_m$.

Phase 2: Let E consist of all remaining position-argument pairs that are potentially effective, *i.e.,* it consists of any (x, a) such that there exists y with $xP^a y$ and such that there is no $i = 1, \ldots, n$ with $(x, a) = (x_i, a_i)$. Index this set as $E = \{(x_{n_1+1}, a_{n_1+1}), \ldots, (x_{n_2}, a_{n_2})\}$, and in rounds $m = n_1 + 1, \ldots, n_2$, we set $x^m = x_m$ and $a^m = a_m$, extending the definition of the protocol to the first n_2 rounds.

Phase 3: In rounds $m = n_2 + 1, \ldots, n_2 + n_1$, we specify that the protocol again run through the elements of the top cycle in the order of their indexing, so $x^m = x_{m-n_2}$ and $a^m = a_{m-n_2}$, extending the definition of $\mathfrak{P}$ to the first $n_1 + n_2$ rounds.

We complete the specification of the protocol $\mathfrak{P}$ by repeating this sequence thereafter with periodicity $n = n_1 + n_2$. Now, consider a myopic discussion $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ for this protocol. Note that $z^2 \in TC$. Indeed, if $z^1 \in TC$, this holds because either $z^2 \neq z^1$, in which case $x^1 P^* z^1$ implies $z^2 = x^1 \in TC$, or $z^2 = z^1$. Then the status quo remains in the top cycle thereafter. Next, we claim that for every multiple of $n = n_1 + n_2$, say αn (where α is a positive integer), the status quo in round $\alpha n + 1$ is x_{n_1}. Indeed, note that in rounds $m = (\alpha - 1)n, \ldots, \alpha n, \alpha n + 1$, the status quo evolves as $z^{(\alpha-1)n}, \ldots, z^{\alpha n+1}$, and specifically, in rounds

$$(\alpha - 1)n + n_1 + 1, \ldots, (\alpha - 1)n + n_2 + 1,$$

the status quo is determined by the potentially effective pairs $(x, a) \in E$ in Phase 2. In the last round of this phase, say, $\ell = (\alpha - 1)n + n_2 + 1$, the status quo belongs to the top cycle, *i.e.,* $z^\ell = x_i$ for some $i = 1, \ldots, n_1$, and at this point, the protocol enters Phase 3 and runs through the top cycle in order of indexing. In the subsequent $i - 1$ rounds,

$$(\alpha - 1)n + n_2 + 2, \ldots, (\alpha - 1)n + n_2 + i,$$

if the status quo changes, then it is replaced by the challenging position, *i.e.,* for all $m = (\alpha - 1)n + n_2 + 1, \ldots, (\alpha - 1)n + n_2 + i - 1$, if $z^{m+1} \neq z^m$, then $z^{m+1} = x^m$. After this, the status quo evolves according to the cycle: if the status quo changes in round $m + 1$, where $m = (\alpha - 1)n + n_2 + j$, then we have $z^{m+1} = x_j$, $z^{m+2} = x_{j+1}, \ldots, z^{\alpha n+1} = x_{n_1}$. And if the status quo does not change during those $i - 1$ rounds, then in round $m = (\alpha - 1)n + n_2 + i$, the status quo $z^m = x_i$ is challenged by position is x_i, and so $z^{m+1} = x_i$, and then we have $z^{m+2} = x_{i+1}, \ldots, z^{\alpha n+1} = x_{n_1}$. In both cases, at the end of Phase 3, the status quo is $z^{\alpha n+1} = x_{n_1}$, as claimed.

To show that $\Lambda(\mathfrak{D}) = TC$ for every myopic discussion consistent with the protocol $\mathfrak{P}$, consider any position $x \in TC$, and let αn be any multiple of $n = n_1 + n_2$. By the above claim, the status quo in round $\alpha n + 1$ is $z^{\alpha n+1} = x_{n_1}$. At this point, the protocol enters Phase 1, and the status quo evolves as $z^{\alpha n+2} = x_1$, $z^{\alpha n+3} = x_2$, $\ldots$, $z^{\alpha n+n_1+1} = x_{n_1}$. Since the positions $x_1, \ldots, x_{n_1}$ exhaust the top cycle, we have $x_i = x$ for some i, and thus $z^{\alpha n+i+1} = x$. Since

α is an arbitrary positive integer, it follows that x belongs to the limit set of $\mathfrak{D}$, as required. □

Theorem 5 implies that the limit set of a myopic discussion can coincide with the top cycle and, thus, be quite large. What is worse is that the top cycle can contain positions that are dominated, so an implication is the possibility that a myopic discussion can visit (and revisit) dominated positions an infinite number of times. Note that the theorem holds for any total p, even under transitivity, and so the next example can use a total, transitive p, illustrating the possibility in a highly structured case.

Example (Myopic Discussion May Visit Dominated Position): Assume there are four positions, $X = \{A, B, C, D\}$, and two arguments, $A = \{a, a'\}$. Define two linear orders, P^a and $P^{a'}$, as follows:

P^a	$P^{a'}$
C	D
D	A
A	B
B	C

and let p be the corresponding set-valued relation: for all $x, y \in X$, $a \in p(x, y)$ if and only if $xP^a y$, and $a' \in p(x, y)$ if and only if $xP^{a'} y$. By Theorem 1, p is total and transitive, and by Theorem 2, D dominates both A and B, because it is ranked higher than both positions by each argument; and both A and D dominate B, because they are ranked higher than B by each argument. Note that $AP^*BP^*CP^*DP^*A$, so the top cycle contains every position. Nevertheless, by Theorem 5, there is a rotating protocol such that every myopic discussion for that protocol cycles through the four positions endlessly. □

5 Constructive Discussion

The analysis of the previous section has revealed that myopic discussion, as a mode of democratic deliberation, fails to meet many ideals of deliberative democracy. Unless there is an unassailable position, myopic discussion is inconclusive and may cycle endlessly through positions, and it necessarily fails to produce unanimous agreement. Furthermore, the long-run behavior of myopic discussion can be too broad ranging and return repeatedly to inferior positions: not only can its limit set be very large, but the limit set of myopic discussion may also contain dominated positions, even when p is total and transitive. Thus, simply adding a process of deliberation does not, *per se,* solve the problem of democratic justification and legitimacy raised by many impossibility theorems of social choice theory. If we wish democratic deliberation to serve as a medium for democratic justification and legitimacy, then we need to impose further structure on democratic deliberation itself.

In this section, we introduce the concept of a "constructive discussion" by imposing further structure on discussion to restrict the evolution of the status quo. A *constructive discussion* is an open discussion $\{(x^m, a^m, z^m)\}$ such that if a position x has previously been justified by an argument a, then a new position y can only be inserted by that same argument if it fares better than x according to a, *i.e.,* $yP^a x$. That is, starting in round m, the new status quo for round $m+1$ is

$$z^{m+1} = \begin{cases} x^m & \text{if } a^m \in p(x^m, z^m), \text{ and for all } k = 1, \ldots, m-1, \\ & x^k = z^{k+1} \text{ and } a^m = a^k \text{ implies } x^m P^{a^m} x^k, \\ z^m & \text{else.} \end{cases}$$

This type of discourse differs from a myopic discussion in that it is context-dependent: the ability to insert a position as status quo by a particular argument depends on the specific history of the discussion, namely, the position must fare better with respect to that argument than any previous status quo that was itself justified by that argument.

When p is transitive, this restriction imposes a sense of directionality on a discussion, and the next result shows the cycles that afflict myopic discussion are impossible in a constructive discussion; in fact, we prove that a constructive discussion must conclude with a position that is maximal for some argument.[5] Moreover, if p is also total, then this concluding position must be undominated, conferring a degree of consensus on the final outcome of discussion.

Theorem 6 (Conclusiveness of Constructive Discussion): Assume X is finite, and p is transitive. Every constructive discussion $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ is conclusive, *i.e.,* there is a position x and some m such that for all $n \geq m$, $z^n = x$. Moreover, there is an argument $a \in A$ such that for all $y \in X$, $a \notin p(y, x)$. Finally, if p is also total, then $x = x^a$, and it is undominated.

Proof: Let $\mathfrak{D} = \{(x^m, a^m, z^m)\}_{m=1}^{\infty}$ be a constructive discussion. To show that $\mathfrak{D}$ is conclusive, suppose toward a contradiction that $z^{m-1} \neq z^m$ for infinitely many m. Let $\{(x^{m_k}, a^{m_k}, z^{m_k})\}_{k=1}^{\infty}$ be a subsequence such that for all k, $z^{m_k - 1} \neq z^{m_k}$. Since A and X are finite, there must be natural numbers $k < \ell$ with $k \geq 2$ such that $(x^{m_k-1}, a^{m_k-1}, z^{m_k-1}) = (x^{m_\ell-1}, a^{m_\ell-1}, z^{m_\ell-1})$. Letting $x = x^{m_k-1} = x^{m_\ell-1}$, we have $x = z^{m_k}$ by the assumption that $z^{m_k-1} \neq z^{m_k}$, and similarly, $x = z^{m_\ell}$. Letting $a = a^{m_k-1} = a^{m_\ell-1}$, it follows that x is re-inserted by argument a after it was previously inserted by the same argument. To see that this is impossible, let T denote the set of rounds t between $m_k - 1$ and $m_\ell - 1$ such that some position x^t is inserted by argument a, *i.e.,*

$$T = \left\{ t \mid m_k - 1 \leq t \leq m_\ell - 1, x^t = z^{t+1} \neq z^t, a = a^t \right\}.$$

We can index this set as $T = \{t_1, \ldots, t_n\}$ so that the indexing of rounds is increasing, *i.e.,*

$$m_k - 1 = t_1 < t_2 < \cdots < t_n = m_\ell - 1.$$

[5] A close reading of the proof of Theorem 6 shows that the conclusiveness of constructive discussion does not rely on the assumption that constructive discussion is open.

Then by definition of constructive discussion, we have

$$x = x^{t_n} P^a x^{t_{n-1}} P^a \cdots x^{t_2} P^a x^{t_1} = x.$$

But transitivity of p implies that P^a is transitive, by Theorem 1, and thus $xP^a x$, contradicting asymmetry of P^a. Thus, $\mathfrak{D}$ is conclusive, and we can let x denote the unique element of the limit set.

To prove the second part of the theorem, suppose toward a contradiction that for every argument $a \in A$, there is a position $y \in X$ such that $a \in p(y, x)$. Since X is finite and x is the conclusion of $\mathfrak{D}$, there exists m such that $z^n = x$ for all $n \geq m$, so that x remains status quo after round m. Letting $a = a^m$, our supposition yields a position y such that $a \in p(y, x)$, but since the discussion is open, there exists $n > m$ such that $(x^n, a^n, z^n) = (y, a, x)$, and then y is inserted by a, *i.e.,* $z^{n+1} = y \neq x$, a contradiction. We conclude that there is an argument a such that for every position y, we have $a \notin p(y, x)$.

Finally, assume p is total, so that P^a is a linear order, by Theorem 1. It follows immediately that x is top ranked in P^a, so that $x = x^a$. Then for every position $y \in X \setminus \{x\}$, we have $a \in p(x, y)$, which is nonempty, and Theorem 2 implies that x is undominated. □

Theorem 6 shows that in our model of constructive discussion, which allows deliberation to continue into the indefinite future, deliberation led by an exogenously enforced protocol will never oscillate but will always conclude with some position that is maximally justified by some argument. In this situation, all other arguments have already been applied with full force, and thus they cannot be used to insert a different position as status quo – so, under the rules of constructive discussion, there is agreement that the construction can proceed no further. Furthermore, when p is total, the position to which a constructive discussion eventually converges is undominated.

As a mode of democratic deliberation, constructive discussion possesses many desirable properties that myopic discussion lacks: it reaches unanimous agreement; the final position to which a constructive discussion converges is maximally justified and, hence, *best* in terms of at least one reason/argument; and a constructive discussion will never conclude with a position that is dominated by another position. Nevertheless, we must ask: How much justification does the process of constructive discussion lend to the final conclusion that is eventually reached?

Theorem 6 ensures that constructive discussions eventually converges to *some* position. However, the result does not ensure that the process of constructive discussion converges to the *same* position for every constructive discussion; rather, it leaves open the possibility that the conclusion can depend on the initial status quo and protocol used. We know that every constructive discussion $\mathfrak{D}$ is conclusive, so we can let $\lambda(\mathfrak{D})$ denote the concluding position of the discussion. Then we define $\Lambda = \{\lambda(\mathfrak{D}) \mid \mathfrak{D} \text{ is a constructive discussion}\}$ as the set of possible conclusions of constructive discussion, and we say constructive discussion is *path dependent* if $|\Lambda| > 1$, *i.e.,* there can potentially be

different conclusions reached by a constructive discussion, so that the conclusion of constructive discussion depends non-trivially on the starting point and protocol.

The next result establishes path dependence of constructive discussion; in fact, we show that if p is total and transitive, then every maximal position x^a can be obtained as the conclusion of constructive discussion.

Theorem 7 (Path Dependence of Constructive Discussion): Assume X is finite and p is total and transitive. For every argument $a \in A$, the position x^a is reached as the conclusion of some constructive discussion, *i.e.,* $\Lambda = \{x^a \mid a \in A\}$. In particular, if there exist $a, a' \in A$ such that $x^a \neq x^{a'}$, then constructive discussion is path dependent.

Proof: Assume p is total and transitive, and consider any argument a and the position x^a, which is top ranked according to the linear order P^a. Let $\tilde{A} \subseteq A$ be the set of arguments with top ranked alternative equal to x^a, *i.e.,* $\tilde{A} = \{\tilde{a} \in A \mid x^{\tilde{a}} = x^a\}$. If there is no argument a' such that $x^{a'} \neq x^a$, then Theorem 5 immediately yields $\lambda(\mathfrak{D}) = x^a$ for every constructive discussion, so henceforth assume that $A \setminus \tilde{A} \neq \emptyset$. Letting $k = |\tilde{A}|$ and $\ell = |A|$, we can index $A \setminus \tilde{A}$ as $A \setminus \tilde{A} = \{a_1, \ldots, a_{\ell-k}\}$ and $\tilde{A}$ as $\tilde{A} = \{a_{\ell-k+1}, \ldots, a_\ell\}$. Furthermore, let E denote the set of potentially effective position-argument pairs (x, a') such that x is not top ranked according to a', and index this set as $E = \{(x_{\ell+1}, a_{\ell+1}), \ldots, (x_n, a_n)\}$.

Define the protocol $\mathfrak{P} = \{(x^m, a^m)\}_{m=1}^{\infty}$ such that for rounds $m = 1, \ldots, \ell$, we have $x^m = x^{a_m}$ and $a^m = a_m$, and for rounds $m = \ell + 1, \ldots, n$, we have $x^m = x_m$ and $a^m = a_m$; and repeat this sequence thereafter. That is, the initial status quo is challenged by argument a_1 and the position that is top ranked for it, then the new status quo is challenged by argument a_2 and the position that is top ranked for it, and so on; in rounds $m = \ell - k + 1, \ldots, \ell$, the position argument pair (x^a, a) challenges the status quo z^m; and in rounds $m = \ell+1, \ldots, n$, the remaining pairs in E appear in the protocol. Let $\mathfrak{D}$ be the constructive discussion for this protocol with initial status quo $z^1 = x^{a_1}$, which means that the first round $m = 1$ trivially determines status quo $z^2 = z^1 = x^{a_1}$ in the second round, and in round $\ell - k + 1$, the status quo does not equal x^a. By Theorem 6, this discussion concludes in a position x that is top ranked by some argument, and we claim that $x = x^a$.

For each argument a', let $m_{a'}$ be the first round in which $x^{a'}$ challenges the status quo, *i.e.,* $m_{a'} = \min\{m \mid x^m = x^{a'}\}$. By construction, $m_{a'} \leq \ell$. Note that prior to round $m_{a'}$, no position has been inserted as status quo by the argument a', and clearly $x^{m_{a'}} = x^{a'} P^{a'} z^{m_{a'}}$, and thus $x^{a'}$ is inserted as status quo in round $m_{a'}$. In particular, x^a challenges the status quo $z^{\ell-k+1}$ in round $\ell - k + 1$, becomes the new status quo in round $\ell - k + 2$, and remains so through round $\ell + 1$, *i.e.,* $z^{\ell-k+2} = \cdots = z^{\ell+1} = x^a$. We must show that no argument $a' \in A \setminus \tilde{A}$ can replace x^a as status quo in rounds $n+1$ onward. For an induction argument, assume x^a remains the status quo until round $m > n$, so that $z^m = x^a$. Clearly, x^a remains the status quo in the next round if $x^m = x^a$,

so assume $(x^m, a^m) = (x, a')$ with $x^m \neq x^a$. In case $a' \in \tilde{A}$, then it is not the case that $xP^{a'}x^{a'} = x^a$, and thus the status quo remains $z^{m+1} = x^a$. In case $a' \in A \setminus \tilde{A}$, note that $x^{a'}$ was previously inserted as status quo by argument a' in round $m_{a'}$, and it is not the case that $xP^{a'}x^{a'}$, and thus the status quo again remains $z^{m+1} = x^a$. We conclude that for all $m > n$, $z^m = x^a$, which implies $\lambda(\mathfrak{D}) = x^a$. Since a was arbitrary, we have proven $\Lambda = \{x^a \mid a \in A\}$, and thus if there are arguments with distinct top ranked positions, then constructive discussion is path dependent. □

Next, we illustrate Theorem 7 in the context of our car purchase example, and we see that each car type can be obtained as the conclusion of a constructive discussion. Thus, while there is agreement that the conclusion of such a discussion cannot be changed, given the history of the discourse, the conclusion is dependent on the particular discussion leading to it, and in this sense, it is *arbitrary*. Put differently, democratic justification in a constructive discussion faces a problem of regress: to fully justify the conclusion reached by some constructive discussion, we would further need to justify the particular history of the constructive discussion leading to the position, as opposed to a different history that would lead to a different conclusion.

Example (Path Dependence of Car Purchases): In the car purchase example, one possible constructive discussion is as follows: beginning with the luxury sedan as status quo, the minivan becomes the new status quo because it is cheaper; the sports car becomes the next status quo because it performs better; and finally, the luxury sedan becomes the status quo because it is more fuel efficient. At this point, the protocol can be specified arbitrarily, because the status quo cannot be changed: cost has already been used to insert the minivan as status quo, and for a car to be inserted again on the basis of cost, it must be cheaper than the minivan, which is impossible; similarly, performance cannot be used again to replace the luxury sedan as status quo. Thus, if the first three rounds of constructive discussion proceed as

$$\begin{array}{rcl}(x^1, a^1, z^1) & = & (V, c, L),\\ (x^2, a^2, z^2) & = & (S, p, V),\\ (x^3, a^3, z^3) & = & (L, f, S),\end{array}$$

then the conclusion of the discussion is the luxury sedan, L. Similarly, if the first three rounds are

$$\begin{array}{rcl}(x^1, a^1, z^1) & = & (S, f, V),\\ (x^2, a^2, z^2) & = & (L, p, S),\\ (x^3, a^3, z^3) & = & (V, c, L),\end{array}$$

then the conclusion of the discussion is the minivan, V. Finally, if the first three rounds are

$$\begin{array}{rcl}(x^1, a^1, z^1) & = & (L, p, S),\\ (x^2, a^2, z^2) & = & (V, c, L),\\ (x^3, a^3, z^3) & = & (S, f, V),\end{array}$$

then the conclusion of the discussion is the sports car, S. Thus, every car that is top ranked according to one of the criteria is a possible conclusion of constructive discussion, *i.e.,* $\Lambda = \{L, V, S\}$, as called for by Theorem 7. □

One important reason why Riker (1982) viewed electoral outcomes as meaningless relates to what he conceived to be the inherent arbitrariness of all electoral outcomes. Even with the same profile of individual preferences, we may reach radically different outcomes depending on the specific voting procedure adopted. Recall that McKelvey's and Schofield's theorems show that under certain circumstances, it is possible to construct a sequence of majority votes that could take us from any status quo alternative to any desired alternative. This is why electoral outcomes, according to Riker, cannot serve as the "true amalgamations of voters' judgments." (Riker 1982: 238)

If it is this arbitrariness of voting outcomes that makes it hard for aggregative voting mechanisms to fully ground democratic justification or legitimacy, then Theorem 7 shows us that democratic deliberation, in the form of constructive discussions, may not take us far. In their 1997 paper, Knight and Johnson conjectured that "there is very good reason to suspect that the outcome of political debate depends heavily upon factors such as the sequence in which participants speak and the point at which debate is terminated" (Knight and Johnson 1997: 291) Theorem 7 shows that Knight and Johnson' speculation is a theoretical possibility in our model of constructive discussion, and it instructs us that we cannot assume that discussion – even in the constructive sense considered in this section – will solve the standard problems of aggregative forms of democracy.

6 A Strategic Model of Debate

There are three main takeaway points from the theory of discussions set forth in the previous sections:

1. Democratic deliberation in the form of myopic discussion is inconclusive unless there is an unassailable position, and otherwise, it is indeterminate and can result in continual disagreement; in fact, there is a protocol that produces the entire top cycle as the limit set of myopic discussion.

2. Constructive discussion resolves this potential for disagreement: it leads, via an agument-climbing dynamic, to convergence at a single position, and this position is maximally justified according to at least one reason/argument.

3. Yet, there is a sense in which a constructive discussion is still unable to confer full democratic justification/legitimacy to its final outcome, as the process of constructive discussion is inherently path dependent: for every conclusion reached by some constructive discussion, there will always be another constructive discussion that will generate a different conclusion.

Now, we should note that there are two important restrictions imposed on constructive discussions. The first is that the whole process of a constructive discussion proceeds via an arbitrary, exogenous protocol; and because the protocol can be specified arbitrarily, we obtain the indeterminacy result of Theorem 7 as a consequence. Second, the model does not allow for conflicting interests of the participants, and thus it is silent on the effects of strategic incentives faced by the participants of the discussion process. One advantage of constructive discussion over myopic discussion, as Theorem 6 shows, is that it eventually reaches unanimous agreement toward a single position. However, a critic might hesitate to endorse this convergence result, as s/he may think that the result is simply the product of these two restrictions imposed on our theory of constructive discussions; once we allow, the critic might claim, deliberation to proceed endogenously among parties who have diametrically opposed interests, convergence or agreement may quickly break down and may lead to conflict or even extreme polarization. (Gutman and Thomson 2004: 54) In this section, we explore this possibility and consider an alternative to constructive discussion, which we call "debate," such that the protocol arises endogenously and is the product of strategic behavior on the part of participants.

We will define a debate as an equilibrium path of play in a particular extensive form game of perfect information that we call the *debate game*.[6] By perfect information, we mean that the participants of the debate game know all past moves taken place in previous rounds – i.e. each participant knows which position has been inserted or retained by which argument in which round, and so forth. Normally, deliberative democratic theorists assume that democratic deliberation embodying idealized deliberative procedure occurs under ideailzed circumstances by ideally rational agents. (Cohen 1997a; Estlund 1997; Habermas 1990; Rawls 1997) It is natural to assume that ideally rational agents would at least remember what positions have been inserted or retained by what arguments in previous rounds of the debate game. Hence, assuming that our debate game is perfect information is consistent with the general setup of democratic deliberation proposed by deliberative democratic theorists.

There are two players, numbered 1 and 2, who alternately argue for different positions, with player 1 moving in all odd rounds, and player 2 moving in all even rounds. An initial status quo $z^1 \in X$ is given, and in any round m, the active player can put forth any position x^m and argue for this using any argument a^m; formally, the action set of a player at any history is $X \times A$. In any round m of the debate, actions $(x^1, a^1), \ldots, (x^m, a^m)$ determine a sequence $(z^2, \ldots, z^{m+1})$ of status quo positions as follows: for each $t = 1, \ldots, m$,

$$z^{t+1} = \begin{cases} x^t & \text{if } a^t \in p(x^t, z^t), \text{ and for all } k = 1, \ldots, t-1, \\ & x^k = z^{k+1} \text{ and } a^t = a^k \text{ implies } x^t P^{a^t} x^k, \\ z^t & \text{else.} \end{cases}$$

[6]For a reference on non-cooperative games and the concepts of Nash equilibrium and subgame perfect equilibrium, see Osborne and Rubinstein (1994).

Note that if $x^1 = z^1$, then the position x^1 is justified as status quo, *i.e.,* $x^1 = z^2$, by any argument. Thus, the evolution of the status quo in the debate game follows the same rule as that of constructive discussion: if a position x has previously been justified by an argument a, then a new position y can only be inserted as status quo by that same argument if it is ranked higher than x according to a, *i.e.,* $yP^a x$.

A *history of length* m, denoted $h^m = ((x^1, a^1, z^1), \ldots, (x^m, a^m, z^m))$, lists arguments made for different positions, along with the sequence of status quo positions determined by the players' actions[7] for the first m rounds, and an *infinite history,* denoted $h^\infty = \{(x^m, a^m, z^m)\}_{m=1}^\infty$ lists positions, arguments, and status quo positions for each round $m = 1, 2, \ldots$

A (pure) *strategy* for player $i = 1, 2$ is a mapping, denoted σ_i, from every history at which the player is active into the set of possible actions. Given an initial status quo z^1, a pair of strategies (σ_1, σ_2) then determines a *path of play,* which consists of the sequence of positions, arguments, and status quo positions along the path of play. Formally, this is the unique infinite history $h^\infty = \{(x^m, a^m, z^m)\}_{m=1}^\infty$ such that for all m, (i) if m is even, then $\sigma_1((x^1, a^1, z^1), \ldots, (x^m, a^m, z^m)) = (x^{m+1}, a^{m+1})$, and (ii) if m is odd, then $\sigma_2((x^1, a^1, z^1), \ldots, (x^m, a^m, z^m)) = (x^{m+1}, a^{m+1})$.

To complete the description of the game, we specify a payoff to each player for each possible infinite history. One standard assumption that is almost universally endorsed by deliberative democratic theorists is that, in modern pluralistic societies, political deliberation occurs under conditions of irreconciliable moral and political disagreement. "A deliberative democracy is a pluralistic association" (Cohen 1997a: 72) explains Cohen, and, according to Gaus, this entails that, even under relatively favorable circumstances, "sincere reasoners will find themselves in principled disagreements." (Gaus 1997: 231) Gutman and Thomson explains that "the general aim of deliberative democracy is to provide the most justifiable conception for dealing with *moral disagreement* in politics." (Gutman and Thomson 2004: 10, emphasis added) Knight and Johnson goes further to claim that the whole purpose of democratic deliberation is to resolve "*political conflict.*" (Knight and Johnson 1994: 285)

We want our model of debate to reflect these conditions of irreconciliable moral and politcal disagreements. To do so, we specify payoffs in the following way. Let $u_i(x)$ denote the payoff to participant i from final position x. To incorporate irreconciliable moral and political disagreement between the two participants, we will assume that the debate game is *competitive* in the sense that for all distinct $x, y \in X$, either $u_1(x) > u_1(y)$ and $u_2(x) < u_2(y)$ hold or the reverse inequalities hold. Thus, moving from one position to another will always generate disagreement – i.e. one participant will favor such a move while the other participant will disfavor such a move. Without loss of generality, index the finite set of positions as $X = \{x_1, \ldots, x_n\}$, and assume the indexing

[7]Technically, our inclusion of the status quo positions z^m for $m \geq 2$ is redundant, as they follow from the initial status quo and actions taken in each round.

is in the order of participant 1's preference: $u_1(x_1) < u_1(x_2) < \cdots < u_1(x_n)$. We consider only pure strategies, so it is again without loss of generality (by a monotonic transformation of payoffs) to assume the game is zero-sum, *i.e.,* for all $x \in X$, we have $u_1(x) + u_2(x) = 0$.

We would like to emphasize that assuming that payoffs are zero-sum does not have any substantive philosophical meaning or normative implications. It does *not*, for instance, mean that the final position reached through the process of debate will necessarily be *partisan* in the sense that it attends to the interests of the 'winners' while ignoring the interests of the 'losers', in which case, our model of debate will go against the general spirit of deliberation sought by deliberative demcoratic theorists who see democratic deliberation as a device for promoting the common good. (Elster 1997; Cohen 1997a, 1997b; Guman and Thomson 2004) What is essential is the existence of *disagreement* in our model of debate, and the assumption that payoffs are zero-sum, in the sense that $u_1(x) + u_2(x) = 0$ for all positions, is merely a technical convenience that is made innocuously without any loss of generality: the payoffs of the two participants of our debate model are ordinal, and so any specification of opposing payoffs can be transformed by a monotonic transformation to satisfy the zero-sum assumption without affecting the subgame perfect equilibria.

Thus, our model of debate is both constructive and competitive in accordance with the deliberative environment assumed by deliberative democratic theorists. Given the payoff functions u_1 and u_2, we assign a payoff to each infinite history h^∞ as follows: as a debate uses the same transition rule for the status quo as constructive discussion, Theorem 3 implies that the history is conclusive, in the sense that there is a position x and round m such that for all $k \geq m$, we have $z^k = x$, and then we specify that player 1's payoff from h^∞ is $u_1(x)$, and player 2's payoff is $u_2(x)$.[8] That is, each player cares only about the conclusion of the debate game, and they evaluate the concluding position according to the payoff functions u_1 and u_2.

We analyze the debate game as a two-player, zero-sum game game of complete information, and we employ concepts of Nash equilibrium and subgame perfect equilibrium to understand the behavior of rational participants in a debate. To apply the former concept, we view the players as choosing strategies σ_1 and σ_2 simultaneously, before the debate game is played; then $\sigma = (\sigma_1, \sigma_2)$ is a Nash equilibrium if neither player $i = 1, 2$ can increase her payoff, *i.e.,* obtain a more desirable conclusion, by unilaterally deviating to another strategy σ_i'. The concept of subgame perfect equilibrium applies this notion at all subgames in the larger extensive form game: after any finite history h^m, we can imagine that the players simultaneously have the opportunity to revise their strategies for the remainder of the game, and a subgame perfect equilibrium is a pair of strategies such that neither player can gain by unilaterally revising her strategy following any such history. Essentially, subgame perfection

[8] As remarked in footnote 4, the proof of Theorem 3 does not use the assumption that constructive discussion is open, and thus it applies equally well to debate.

removes the possibility of "non-credible" threats that could conceivably play a role in Nash equilibria.

A *debate* is any path of play $h^\infty = \{(x^m, a^m, z^m)\}_{m=1}^\infty$ generated by a Nash equilibrium of the debate game. We establish that the debate game possesses a subgame perfect equilibrium, and since every subgame perfect equilibrium is Nash, the existence of a debate follows. We use notation and results from the previous section; in particular, we denote a debate by $\mathfrak{D}$, and we have already observed that every debate is conclusive, so that the limit set $\Lambda(\mathfrak{D})$ of a debate is a singleton.

Since constructive discussion, in general, is path dependent, we are now interested in whether a debate, which endogenizes the protocol governing the discussion, similarly suffers from the problem of path dependence, or whether strategic incentives of the players isolate a unique position that does not depend on the initial status quo. The debate game does generally possess multiple Nash equilibria, but our equilibrium characterization establishes that all Nash equilibria determine the same outcome, and that this position is *independent* of the initial status quo: the unique conclusion of debate is a compromise position that can be identified by the primitives of the model. In what follows, we let $\Lambda^* = \{\lambda(\mathfrak{D}) \mid \mathfrak{D} \text{ is a debate}\}$ denote the set of possible conclusions of debate.

Our characterization of Nash equilibria of the debate game rests on the idea of a compromise position. When the number $|A|$ of arguments is odd, we define the *compromise position* x^* as the unique position such that x^* is top ranked for some argument, the number of arguments with a top ranked position better than x^* for player 1 is less than $|A|/2$, and the number of arguments with a top ranked position better than x^* for player 2 is also less than $|A|/2$. Formally, $x^* = x^a$ for some $a \in A$, and

$$\sum_{a \in A} I_{a,1}(x^*) \quad \leq \quad \frac{|A|}{2} \tag{1}$$

and

$$\sum_{a \in A} I_{a,2}(x^*) \quad \leq \quad \frac{|A|}{2}, \tag{2}$$

where $I_{a,i}(x)$ is an indicator equal to one if $u_i(x^a) > u_i(x)$ and equal to zero otherwise. When $|A|$ is even, there may be one or two positions that are top ranked by different arguments and satisfy both (1) and (2). We say x^* is the *compromise position* if it is the unique position satisfying the inequalities, or if there are two such positions, then it is the one preferred by player 1; formally, letting x_k and x_ℓ solve (1) and (2) with $k < \ell$, we define $x^* = x_\ell$.

The next theorem establishes that when p is total and transitive, the unique Nash equilibrium outcome of the debate game is the compromise position, regardless of the initial status quo. An immediate implication is that debate,

in contrast to constructive discussion, does not suffer from the problem of path dependence.[9]

Theorem 5 (Path Independence of Debate): Assume X is finite, and p is total and transitive. Then there is at least one debate, and the conclusion of every debate is the compromise position: $\Lambda^* = \{x^*\}$.[10]

The next example provides an extended illustration of the idea of debate and the result of Theorem 5 in the context of the car purchase example, in which we imagine a husband and wife debating about which among the three possible cars to buy as a family car.

Example (Constructive Debate about Car Purchase): In the car purchase example, let the wife be player 1 and the husband be player 2, and assume $u_1(S) > u_1(L) > u_1(V)$, while $u_2(S) < u_2(L) < u_2(V)$. Note that the top ranked positions of the arguments are $x^f = L$, $x^c = V$, and $x^p = S$. Because there are three arguments in this example, the compromise position is uniquely defined by (1) and (2), and it is just $x^* = L$. By Theorem 5, the unique conclusion of the debate between the husband and wife is thus the luxury sedan, regardless of the initial status quo. The theorem does not state that there is a unique debate, but the construction used in the proof provides one debate in this example. Given any status quo z^1, the path of play of the constructed equilibrium is as follows: the wife moves first in round 1 with $(x^1, a^1) = (V, c)$, which changes the status quo in round 2 to $z^2 = V$; then the husband moves in round 2 with $(x^2, a^2) = (S, p)$, which changes the status quo in round 3 to $z^3 = S$; and the wife moves in round 3 with $(x^3, a^3) = (L, f)$, after which the status quo remains $z^m = L$ for all future rounds $m \geq 4$.

Here is an intuitive story that describes what is happening on the path of play. Let the initial status quo be any position, say, the luxury sedan, so that $z^1 = L$. First, it is the wife's turn to argue for a position. The wife's ideal outcome is the sports car. However, she knows that if she inserts the sports car on the basis of performance now, *i.e.,* if she plays $(x^1, a^1) = (S, p)$, then this creates an opening for her husband: he can eliminate the sports car from the debate by inserting the luxury sedan on the basis of fuel economy in

[9] In the Appendix, we discuss a different approach to modeling strategic interaction among participants that extends the Bipartisan set of Laffond, Laslier, and Le Breton (1993, 1997). In this approach, we assume participants simultaneously choose positions in a strategic form game, and a participant cares only about the net number of arguments in favor of her position.

[10] It has been shown in the formal literature of deliberation that the protocol or post-deliberation voting rule together with various forms of uncertainty can generate strategic incentives and can create path dependencies and steer away from consensus toward a uniquely determined final outcome.(Austen-Smith and Feddersen 2006; Ottaviani and Sorensen 2001) Theorem 5 does not necessarily contradict these prior results as the strategic incentives faced by the debaters in our model are different; in our model, the debaters primarily care about reaching a mutual consensus toward their favored position rather than reaching an independently correct verdict (Austen-Smith and Feddersen 2006) or improving their reputation to appear as experts (Ottaviani and Sorensen 2001).

round 2, which will eventually lead the debate to conclude with the minivan, which is her worst position. So, the wife, instead, preempts this by proposing the minivan on the basis of cost herself by saying, "Why don't we consider the minivan? It's the cheapest among the three and you seem to like it very much." That is, she plays $(x^1, a^1) = (V, c)$, and as a consequence, the status quo changes to the minivan: $z^2 = V$.

Now, it is the husband's turn to argue for a position. The current status quo is the minivan, his favorite position, and hence, the husband would want to preserve it as the final outcome, if possible. He could maintain it for the current period by responding, for example, "Yes, you're right. It's a really good price," which corresponds to formally proposing $(x^2, a^2) = (V, c)$. However, if the husband does that, then this creates an opening for his wife: she can eliminate the minivan with the luxury sedan on the basis of fuel economy in round 3, which (for similar reasons as before) will eventually lead the debate to conclude with the sports car, which is his worst position. So, the husband preempts this by proposing the sports car on the basis of performance himself, saying, "Actually, how about we consider the sports car? It has the best performance and you seem to really like it," which corresponds to $(x^2, a^2) = (S, p)$. As a consequence the status quo changes to the sports car: $z^3 = S$.

Finally, it remains for the wife to argue for a position. The current status quo is the sports car, her favorite position, but there is no way for her to maintain it, as the only argument that has not been used to its full extent is fuel economy, according to which the luxury sedan is ranked first. And, since the husband strictly prefers the luxury sedan to the sports car, he will surely insert the luxury sedan on the basis of fuel economy sooner or later if the wife does not do so herself. Hence, the debate will eventually conclude with the luxury sedan, which is the car that is maximally justified by the fuel economy argument and is, consistent with Theorem 5, the compromise position in this example. □

Several points are noteworthy, both technical and philosophical. First, just like constructive discussions, every debate is conclusive, avoiding the cycling problems that afflict myopic discussions. Second, the *unique* conclusion of every debate is the compromise position, independent of the initial status quo; thus, in contrast to constructive discussion, debate is *path-independent.* Third, when the total number of positions is odd, neither player has an advantage in debate, as the two make symmetric compromises at position x^*, the middle-ranked position of each player. When the total number of positions is even, player 1 has an advantage, but only a slight one. This means that the procedure of a constructive debate itself incorporates *fairness,* in the sense that everybody in a constructive debate is situated roughly equally, and no player has a significant advantage in their ability to effect a final position in their favor. Fourth, the incentives of constructive yet competitive debate do lead to an outcome that is maximally justifiable with respect to some argument, consistent with the argument climbing properties of constructive discussion.

This means that the compromise position is not simply a "compromise," in the sense that everybody settles with something that meets some low bar of mutual acceptability, but rather it is a position that is regarded as *best* by both parties with respect to at least one argument.

Lastly, note that the strategic incentives faced by the participants in our model of debate are perfectly consistent with the general framework of deliberative democratic theory. Although the normative conditions of deliberative democratic theory preclude purely selfish or self-centered considerations, different people are still allowed to form different policy preferences (concerning which policy would be best for the group) on the basis of their different conceptions of the public or common good. As already noted, democratic pluralism is an assumption that is universally endorsed by deliberative democratic theorists in general. And, even if the debate participants are assumed to respond solely to what Harbermas calls the "force of better argument" (Habermas 1975: 108), this does not rid democratic deliberation of all strategic considerations, for in democratic deliberation, each participant must consider how to best utilize the set of common reasons accepted by all to construct arguments that would persuade other similarly-motivated-but-politically-divergent participants to reach an agreement on a policy position that she sincerely believes to be best – not just for her, but for society as a whole.[11]

7 Conclusion

No democratic theorist presumes that successful democratic deliberation can happen automatically. This is why so many deliberative democratic theorists have strived to clearly define the institutional requirements and rules of ideal deliberative procedure. (Cohen 1997a, 1997b; Estlund 1997; Gutman and Thomson 2004; Habermas 1990) Usually, deliberative democratic theorists have emphasized that ideal deliberative procedure should include: (a) both formal and substantive *equality* among the participants of deliberation, (b) *freedom* to express one's opinions and propose new positions, (c) *fair equal opportunity* to speak and participate in deliberation, and (d) *reciprocity* in the sense that positions should be supported by reasons that all endorese. In this way, "the ideal deliberative procedure is meant to provide a model for institutions to mirror." (Cohen 1997a: 73) The thought is that when real-world democratic institutions approximate the institutional requirements of ideal deliberative procedure, this will lead society to arrive at democratic decisions that are fully justified by reasoned agreement through democratic deliberation. However, for most deliberative democratic theorists, whether approximating ideal deliberative procedure would really be enough to arrive at democratically

[11]For illustrative purposes, we have used an example of a husband and wife debating about which car to buy, but the point of the debate was to collectively decide on which car would be best for *the whole family*, not just for him or her.

legitimate decisions remains an educated guess. This is the part where we believe formal analysis can help and empower normative philosophical theorizing.

In this paper, we have seen, through the lens of formal analysis, that not only can democratic deliberation take many different forms, but different forms of democratic deliberation can either fail or succeed to confer democratic legitimacy in ways that were not entirely apparent prior to conducting such formal analyses. All three modes of democratic deliberation that we have discussed in the paper – namely, myopic discussion, constructive discussion, and constructive debate – are consistent with the general characterizations of ideal deliberative procedure laid forth by deliberative democratic theorists. However, the three modes of democratic deliberation differ in the extent to which they confer democratic legitimacy to their final outcomes or the lack thereof. Unless there is an unassailable position, myopic discussion is inconclusive and can result in continual disagreement. Constructive discussion resolves this issue and converges to a single position via an argument-climbing dynamic, but is inherently path dependent, and, hence, arbitrary. In contrast to these two other modes of democratic deliberation, our model of debate generates an outcome that is unique, path-independent, and best with respect to at least one reason or argument. Moreover, this outcome represents a compromise among the participants, and thus it has some justification in terms of fairness. For these reasons, it might be said that, among the three modes of democratic deliberation considered in this paper, debate confers better democratic justification or legitimacy to its resulting outcomes than constructive or myopic discussions. Far from generating conflict and extreme polarization, it was the very addition of diametrically opposed disagreements, along with the endogenous formation of the agenda by strategic players, that led to the greater degree of legitimacy under debate.

References

[1] Arendt, Hannah (1973), *On Revolution*, Harmondsworth: Pelican Books

[2] Arrow, Kenneth (1951/1963), *Social Choice and Individual Values (2nd Edition)*, Yale University Press

[3] Bohman, James and William Rehg (1997), *Deliberative Democracy: Essays on Reason and Politics*, MIT Press

[4] Christiano, Thomas (1997), "The Significance of Public Deliberation" contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 243–278

[5] Chung, Hun (2017), "The Instability of John Rawls's "Stability for the Right Reasons," *Episteme*, published online (doi:10.1017/epi.2017.14)

[6] Cohen, Joshua (1997a), "Deliberation and Democratic Legitimacy," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 67–92

[7] Cohen, Joshua (1997b), "Procedure and Substance in Deliberative Democracy," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 407–438

[8] Coleman, Jules and John Ferejohn (1986), "Democracy and Social Choice," *Ethics* 97 (1), 6–25

[9] Dietrich, Franz and Christian List (2013), "A Reason-Based Theory of Rational Choice," *Noûs* 47 (1), 104–134

[10] Dryzek, John and Christian List (2003), "Social Choice Theory and Deliberative Democracy: A Reconciliation," *British Journal of Political Science* 33, 1–28

[11] Duggan, John and Thomas Schwartz (2000), "Strategic Manipulability without Resoluteness or Shared Beliefs: Gibbard-Satterthwaite Generalized," *Social Choice and Welfare,* 17, 85–93

[12] Elster, Jon (1997), "The Market and the Forum: Three Varieties of Political Theory," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 3–34

[13] Estlund, David (1997), "Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 173–204

[14] Gaus, Gerald (1997), "Reason, Justification, and Consensus: Why Democracy Can't Have It All," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 205–242

[15] Gaus, Gerald (2016), *The Tyranny of the Ideal: Justice in a Diverse Society*, Princeton University Press

[16] Gibbard, Allan (1973), "Manipulation of Voting Schemes: A General Result," *Econometrica* 41 (4), 587–601

[17] Gutman, Amy and Dennis Thomson (2004), *Why Deliberative Democracy?*, Princeton University Press

[18] Habermas, Jürgen (1975), *Legitimation Crisis* (translated by Thomas McCarthy), Boston: Beacon Press; London: Heinemann

[19] Habermas, Jürgen(1990), "Discourse Ethics: Notes on a Program of Philosophical Justification," contained in Jürgen Habermas, *Moral Consciousness and Communicative Action* (translated by C. Lenhardt and S. W. Cicholsen), MIT Press, 43–115

[20] Habermas, Jürgen(1997), "Popular Sovereignty as Procedure," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 35–66

[21] Hong, Lu and Scott Page (2004), "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers," *Proceedings of the National Academy of Sciences* 101 (46), 16385–16389

[22] Ingham, Sean (2012), "Disagreement and Epistemic Arguments for Democracy," *Politics, Philosophy & Economics* 12 (2), 136–155

[23] Kogelman, Brian and Stephen Stich (2016), "When Public Reason Fails Us: Convergence Discourse as Blood Oath," *American Political Science Revew* 110 (4), 717–730

[24] Knight, Jack and James Johnson (1994), "Aggregation and Deliberation: On the Possibility of Democratic Legitimacy," *Political Theory* 22 (2), 277–296

[25] Knight, Jack and James Johnson (1997), "What Sort of Equality Does Deliberative Democracy Require?", contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 279–320

[26] Knight, Jack and James Johnson (2007), "The Priority of Democracy: A Pragmatist Approach to Political-Economic Institutions and the Burdern of Justification," *American Political Science Review* 101 (1), 47–61

[27] Landa, Dimitri and Adam Meirowitz (2009), "Game Theory, Information, and Deliberative Democracy," *American Journal of Political Science* 53 (2), 427–444

[28] Landemore, Helene (2013), *Democratic Reason,* Princeton University Press

[29] List, Christian and Robert Goodin (2001), "Epistemic Democracy: Generalizing the Condorcet Jury Theorem," *The Journal of Political Philosophy* 9 (3), 277–306

[30] List, Christian, Robert Luskin, James Fishkin, and Lain McLean (2013), "Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls," *The Journal of Politics* 75 (1), 80–95

[31] Mackie, Gerry (2003), *Democracy Defended,* New York, NY: Cambridge University Press

[32] May, Kenneth (1952), "A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision," *Econometrica* 20, 680–684

[33] McKelvey, Richard (1976), "Intransitivities in Multi-dimensional Voting Models and Some Implications for Agenda Control," *Journal of Economic Theory* 12, 472–482

[34] McKelvey, Richard (1979) "General Conditions for Global Intransitivites in Formal Voting Models," *Econometrica* 47, 1085–1112

[35] Osborne, Martin and Ariel Rubinstein (1994) *A Course in Game Theory,* MIT Press

[36] Patty, John (2008), "Arguments-Based Collective Choice," *Journal of Theoretical Politics*" 20 (4), 379-414

[37] Patty, John and Elizabeth Maggie Penn (2011), "A Social Choice Theory of Legitimacy," *Social Choice and Welfare,* 36, 365–382

[38] Patty, John and Elizabeth Maggie Penn (2014), *Social Choice and Legitimacy: The Possibilities of Impossibility,* Cambridge University Press

[39] Perote-Peña, Juan and Ashley Piggens (2015), "A Model of Deliberative and Aggregative Democracy," *Economics and Philosophy* 31, 93–121

[40] Plott, Charles (1967), "A Notion of Equilibrium and Its Possibility Under Majority Rule," *American Economic Review* 57, 787–806

[41] Rawls, John (1971/1999). *A Theory of Justice (revised edition),* Harvard University Press

[42] Rawls, John (1993/2005), *Political Liberalism (expanded edition),* Columbia University Press

[43] Rawls, John (1997), "The Ideal of Public Reason," contained in *Deliberative Democracy: Essays on Reason and Politics* (edited by James Bohman and William Rehg), MIT Press, 93–141

[44] Rawls, John (2001), *Justice as Fairness: A Restatement,* Belknap Harvard

[45] Riker, William (1982), *Liberalism Against Populism,* Waveland Press

[46] Satterthwaite, Mark A. (1975), "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory* 10 (2), 187–217

[47] Schofield, Norman (1978), "Instability of Simple Dynamic Games," *Review of Economics Studies* 45, 575–594

[48] Schumpeter, Joseph (2003), *Capitalism, Socialism and Democracy,* Routledge (London, New York): Taylor and Francis e-Library

[49] Shapiro, Ian (2003), *The State of Democratic Theory,* Princeton University Press

Appendix: Proof of Theorem 8

The proof shows, by induction, that the compromise position x^* is the unique subgame perfect equilibrium outcome. An immediate implication is that the compromise position is a Nash equilibrium outcome. Moreover, because the debate game is a two-player, zero-sum game, it follows that equilibrium payoffs are unique: the value of the game for player 1 is $u_1(x^*)$, and the value of the game for player 2 is $u_2(x^*) = -u_1(x^*)$. Every Nash equilibrium gives the players their values, and the only position that gives the value of the game to each player is x^*, and we conclude that the compromise position is the unique Nash equilibrium outcome.

Given any history h^{m-1} of the extensive form of this game, some status quo z^m is determined, and we can consider the strategic form of the subgame at history h^m. Again, this will be a two-player, zero-sum (non-symmetric) game. We solve, recursively, for the equilibrium outcomes of these subgames. In such a subgame, for each argument $a \in A$, we can identify the set $X^a(h^{m-1}, z^m)$ of positions that can be inserted by argument a over each position previously justified by a as

$$X^a(h^{m-1}, z^m) \;=\; \left\{ x \in X \mid \begin{array}{c} \text{for all } k = 1, \ldots, m-1,\ x^k = z^{k+1} \\ \text{and } a^k = a \text{ implies } z^m \neq x P^a x^k \end{array} \right\}.$$

At the initial history h^0, before any actions have been taken, the condition defining this set is vacuously satisfied, so that $X^a(h^0, z^1) = X$. In general, the set $X^a(h^{m-1}, z^m)$ may be empty for some histories. Recall that if $x^{m-1} = z^{m-1}$, then $x^{m-1} = z^m$ is justified as status quo by a^{m-1}, and thus it follows from the above definition that $x^{m-1} \notin X^{a^{m-1}}(h^{m-1}, z^m)$, as the position cannot be re-inserted by the same argument.

Let $A(h^{m-1}, z^m)$ be the set of *active arguments* for which the set of insertable positions is nonempty, *i.e.,*

$$A(h^{m-1}, z^m) \;=\; \{a \in A \mid X^a(h^{m-1}, z^m) \neq \emptyset\},$$

and let $\alpha(h^{m-1}, z^m) = |A(h^{m-1}, z^m)|$ be the number of such arguments. Let $X^*(h^{m-1}, z^m) = \{x^a \in X \mid a \in A(h^{m-1}, z^m)\}$ consist of positions top ranked for at least one argument in $A(h^{m-1}, z^m)$. For future use, observe that if x^{m-1} is top ranked by argument a^{m-1}, then $a^{m-1} \notin A(h^{m-1}, z^m)$. To see this, suppose toward a contradiction that $a^{m-1} \in A(h^{m-1}, z^m)$. In case $x^{m-1} \neq z^m$, then position x^{m-1} is inserted as status quo by argument a^{m-1}, and it cannot be re-inserted by the same argument, which implies $x^{m-1} \notin X^{a^{m-1}}(h^{m-1}, z^m)$. In case $x^{m-1} = z^m$, we have already noted $x^{m-1} \notin X^{a^{m-1}}(h^{m-1}, z^m)$. In either case, the position x^{m-1} cannot be inserted again using a^{m-1}, and since x^{m-1} is top ranked according to a^{m-1}, it follows that no other position can be inserted using the argument. Thus, $X^{a^{m-1}}(h^{m-1}, z^m) = \emptyset$, as required.

Next, we define a notion of *compromise position at* (h^{m-1}, z^m). If the number $\alpha(h^{m-1}, z^m)$ of active arguments is odd, then define the compromise

position $x^*(h^{m-1}, z^m)$ at (h^{m-1}, z^m) as the unique solution $x \in X^*(h^{m-1}, z^m)$ satisfying the inequalities

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,1}(x) \;\leq\; \frac{\alpha(h^{m-1}, z^m)}{2} \tag{3}$$

and

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,2}(x) \;\leq\; \frac{\alpha(h^{m-1}, z^m)}{2}. \tag{4}$$

When $\alpha(h^{m-1}, z^m)$ is even, there may be one or two positions in $X^*(h^{m-1}, z^m)$ satisfying both (3) and (4). We set $x^*(h^{m-1}, z^m)$ equal to the unique position $x \in X^*(h^{m-1}, z^m)$ satisfying (3) and (4), or, if there are two such positions, say x_k and x_ℓ with $k < \ell$, the compromise position is defined by two cases: in case m is odd, then set $x^*(h^{m-1}, z^m) = x_\ell$, and in case m is even, then set $x^*(h^{m-1}, z^m) = x_k$.

The proof of the theorem follows from an induction argument on the number $\alpha(h^{m-1}, z^m)$ of active arguments. First, consider any (h^{m-1}, z^m) such that $\alpha(h^{m-1}, z^m) = 1$, and let a be the argument such that $A(h^{m-1}, z^m) = \{a\}$. In this case, we claim that the compromise position is $x^*(h^{m-1}, z^m) = x^a$. Indeed, if $z^m = x^a$, then it is not possible to change the status quo, and the debate ends with the compromise outcome. Otherwise, $z^m \neq x^a$. Suppose that the equilibrium outcome from (h^{m-1}, z^m) is $x \neq x^*$. For some player i, we have $u_i(x^*) > u_i(x)$, but then i can insert position x^a with argument a to obtain that as the final outcome, a contradiction. Thus, the compromise position x^a is the unique Nash equilibrium outcome, as claimed.

Next, assume the claim is true when the number of active arguments is $1, 2, \ldots, k$. Formally, assume that for all (h^{m-1}, z^m) with $|A(h^{m-1}, z^m)| \leq k$, the unique Nash equilibrium outcome at this subgame is the compromise position $x^*(h^{m-1}, z^m)$. For the induction argument, consider any (h^{m-1}, z^m) with $|A(h^{m-1}, z^m)| = k + 1$. We prove that the unique Nash equilibrium outcome is the compromise position by considering four cases.

Case 1: $|A(h^{m-1}, z^m)|$ is odd, and m is odd. Then player 1 moves. Let $\underline{x}$ minimize u_1 over $X^*(h^{m-1}, z^m)$, let $\underline{a} \in A(h^{m-1}, z^m)$ satisfy $\underline{x} = x^{\underline{a}}$, and suppose that player 1 justifies $\underline{x}$ by argument $\underline{a}$. Let $h^m = (h^{m-1}, \underline{x}, \underline{a})$ be the resulting history, and note that the status quo becomes $z^{m+1} = \underline{x}$. Then by our observation above, the set of active arguments becomes

$$A(h^m, z^{m+1}) \;=\; A(h^{m-1}, z^m) \setminus \{\underline{a}\}.$$

Since the number of active arguments has decreased, the induction hypothesis implies that the unique equilibrium outcome is the compromise position at (h^m, z^{m+1}), say x^*, and this satisfies

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,1}(x^*) \;\leq\; \frac{\alpha(h^m, z^{m+1})}{2} \tag{5}$$

and

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) \;\leq\; \frac{\alpha(h^m, z^{m+1})}{2}. \tag{6}$$

Since $u_1(x^*) \geq u_1(\underline{x})$, we have

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,1}(x^*) \;=\; \sum_{a \in A(h^{m-1}, z^m)} I_{a,1}(x^*),$$

and since

$$\frac{\alpha(h^m, z^{m+1})}{2} \;=\; \frac{\alpha(h^{m-1}, z^m)}{2} - \frac{1}{2}, \tag{7}$$

it follows that x^* satisfies (3).

Moreover, $\alpha(h^m, z^{m+1})$ is even, and thus there may be one or two positions satisfying (5) and (6). If there is just one, then it must appear at the top of at least two arguments, and thus, the inequality in (6) holds strictly at x^*, *i.e.,*

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) \;<\; \frac{\alpha(h^m, z^{m+1})}{2}. \tag{8}$$

If there are two positions satisfying (5) and (6), then the compromise position is the position preferred by player 2, and again the above inequality holds strictly. Since $u_2(\underline{x}) \geq u_2(x^*)$, we have

$$\begin{aligned}
\sum_{a \in A(h^{m-1}, z^m)} I_{a,2}(x^*) \;&\leq\; 1 + \left(\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) \right) \\
&<\; 1 + \frac{\alpha(h^m, z^{m+1})}{2} \\
&=\; \frac{\alpha(h^{m-1}, z^m)}{2} + \frac{1}{2},
\end{aligned}$$

where the second inequality follows from (8), and the equality follows from (7), and therefore x^* satisfies (4).

We conclude that x^* is in fact the compromise position at (h^{m-1}, z^m). We have shown that at history h^m, if player 1 inserts $\underline{x}$ by argument $\underline{a}$, then her payoff in the continuation of the game is $u_1(x^*)$. In equilibrium at the subgame following h^{m-1}, player 1's actions are optimal, and thus her equilibrium payoff in the subgame is at least equal to $u_1(x^*)$. It remains to be shown that player 1 cannot obtain a higher payoff by maintaining the status quo or inserting a different position by another argument. Suppose that player 1 maintains the status quo, so that $z^{m+1} = z^m$, and thus $A(h^m, z^{m+1}) = A(h^{m-1}, z^m)$ and $X^*(h^m, z^{m+1}) = X^*(h^{m-1}, z^m)$. Continuation play then determines an outcome x' following this history h^m. Let $\overline{x}$ minimize u_2 over $X^*(h^m, z^{m+1})$,

and let $\overline{a} \in A(h^m, z^{m+1})$ satisfy $\overline{x} = x^{\overline{a}}$. Then by a symmetric argument, player 2 can insert $\overline{x}$ by argument $\overline{a}$ and obtain the compromise position x^*. Since player 2's equilibrium strategy is optimal at h^m, it follows that $u_2(x') \geq u_2(x^*)$, and we deduce that $u_1(x') \leq u_1(x^*)$. Thus, player 1 cannot obtain a better outcome than x^* by maintaining the status quo.

Now, suppose that at h^{m-1}, player 1 inserts a different position $\tilde{x}$ by an argument $\tilde{a}$, and let $\tilde{h}^m$ be the resulting history with status quo $\tilde{z}^{m+1} = \tilde{x}$. Then the set of active arguments becomes

$$A(\tilde{h}^m, \tilde{z}^{m+1}) \;=\; A(h^{m-1}, z^m) \setminus \{\tilde{a}\},$$

and the positions top ranked for some active argument make up the set

$$X^*(\tilde{h}^m, \tilde{z}^{m+1}) \;=\; X^*(h^{m-1}, z^m) \setminus \{\tilde{x}\}.$$

Since the number of active arguments has decreased, the induction hypothesis implies that the equilibrium outcome is the compromise position at $(\tilde{h}^m, \tilde{z}^{m+1})$, say $\tilde{x}^*$, and this satisfies

$$\sum_{a \in A(\tilde{h}^m, \tilde{z}^{m+1})} I_{a,1}(\tilde{x}^*) \;\leq\; \frac{\alpha(\tilde{h}^m, \tilde{z}^{m+1})}{2} \tag{9}$$

and

$$\sum_{a \in A(\tilde{h}^m, \tilde{z}^{m+1})} I_{a,2}(\tilde{x}^*) \;\leq\; \frac{\alpha(\tilde{h}^m, \tilde{z}^{m+1})}{2}. \tag{10}$$

If $u_1(\tilde{x}) < u_1(x^*)$, then x^* satisfies (9) and (10), and it follows that the compromise position at $(\tilde{h}^m, \tilde{z}^{m+1})$ is equal to the compromise position at (h^m, z^{m+1}), *i.e.*, $x^* = \tilde{x}^*$. Thus, player 1 does not obtain a higher payoff than $u_1(x^*)$.

The remaining case is $u_1(\tilde{x}) > u_1(x^*)$. Then comparing the left-hand sides of (5) and (9) evaluated at x^*, we have

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,1}(x^*) \;\geq\; \sum_{a \in A(\tilde{h}^m, \tilde{z}^{m+1})} I_{a,1}(x^*),$$

and comparing the left-hand sides of (6) and (10), we have

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) \;\leq\; \sum_{a \in A(\tilde{h}^m, \tilde{z}^{m+1})} I_{a,2}(x^*).$$

It follows that for every solution x to (9) and (10), we have $u_1(x) \leq u_1(x^*)$, and in particular, $u_1(x^*) \geq u_1(\tilde{x}^*)$. Again, we conclude that player 1 cannot obtain a payoff higher than $u_1(x^*)$. Therefore, for a history h^{m-1} and status quo z^m in case 1, the unique equilibrium outcome is the compromise position at (h^{m-1}, z^m), namely, x^*.

Case 2: $|A(h^{m-1}, z^m)|$ is odd, and m is even. This case is symmetric to Case 1, interchanging the roles of players 1 and 2.

Case 3: $|A(h^{m-1}, z^m)|$ is even, and m is odd. The argument in this case is similar to that for Case 1. Again, player 1 moves. Define $\underline{x}$ and $\underline{a}$ as in Case 1. If player 1 inserts $\underline{x}$ by argument $\underline{a}$, then again the induction hypothesis is applied, with the implication that the unique equilibrium at (h^m, z^{m+1}) is the compromise position x^*, now the unique solution to (5) and (6), since $\alpha(h^m, z^{m+1})$ is odd. Since $u_1(x^*) \geq u_1(\underline{x})$, we again have

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,1}(x^*) \;=\; \sum_{a \in A(h^{m-1}, z^m)} I_{a,1}(x^*),$$

and again

$$\frac{\alpha(h^m, z^{m+1})}{2} \;=\; \frac{\alpha(h^{m-1}, z^m)}{2} - \frac{1}{2},$$

which implies that x^* satisfies (3). Since $u_2(\underline{x}) \geq u_2(x^*)$, we have

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,2}(x^*) \;=\; 1 + \left(\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) \right),$$

and it follows that

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,2}(x^*) \;<\; \frac{\alpha(h^{m-1}, z^m)}{2} + \frac{1}{2},$$

and thus x^* satisfies (4).

There may be one or two positions satisfying (3) and (4). If there is just one, then we have shown that x^* is equal to the compromise position at (h^{m-1}, z^m). If there are two, say x^* and $\hat{x}$, then we must show that $u_1(x^*) > u_1(\hat{x})$. Otherwise, $x^* < \hat{x}$, and the inequality in (3) must hold with equality at x^*, *i.e.,*

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,1}(x^*) \;=\; \frac{\alpha(h^{m-1}, z^m)}{2}.$$

But this implies

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,1}(x^*) \;=\; \frac{\alpha(h^m, z^{m+1})}{2} + \frac{1}{2},$$

contradicting the fact that x^* is the compromise position at (h^m, z^{m+1}). Thus, $u_1(x^*) > u_1(\hat{x})$, as desired.

We conclude that x^* is, in fact, the compromise position at (h^{m-1}, z^m). We must then show that player 1 cannot obtain a payoff higher than $u_1(x^*)$

by maintaining the status quo or inserting a different position by another argument. Paralleling the argument for Case 1, if player 1 maintains the status quo, then by inserting $\overline{x}$ by $\overline{a}$, player 2 can obtain x^*, and it follows that player 1 cannot be better off as a result. If player 1 inserts a different $\tilde{x}$ by argument $\tilde{a}$, then either $u_1(\tilde{x}) < u_1(x^*)$, in which case the equilibrium outcome by the induction hypothesis remains x^*; or $u_1(\tilde{x}) > u_1(x^*)$, in which case the resulting equilibrium outcome, $\tilde{x}^*$ is equal or less than x^*, and again player 1 cannot obtain a payoff higher than $u_1(x^*)$.

Case 4: $|A(h^{m-1}, z^m)|$ is even, and m is even. This case is symmetric to Case 3, interchanging the roles of players 1 and 2. □