doi:10.1017/S0003055419000674

# A Formal Theory of Democratic Deliberation

HUN CHUNG *Waseda University*

JOHN DUGGAN *University of Rochester*

*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 article helps fill that gap by offering a formal theory of three different modes of democratic deliberation: myopic discussion, constructive discussion, and debate. We show that myopic discussion suffers from indeterminacy of long run outcomes, while constructive discussion and debate are conclusive. 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.*

## 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 dictatorship or 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 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 to the collective decision at hand. It has been known since the work of Marquis de Condorcet in the eighteenth 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 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. In such situations, there is no *Condorcet winner*—an alternative that beats all others in pairwise majority votes—and thus no social alternative that can be unambiguously regarded as the "best" social choice on majoritarian grounds.

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 and 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 typically 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 manipulate 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 Schwartz (2000) have shown that the impossibility result extends even when social ties are

Hun Chung, Associate Professor, Faculty of Political Science and Economics, Waseda University, hunchung1980@gmail.com.

John Duggan, Professor of Political Science and Economics, University of Rochester, dugg@ur.rochester.edu.

Hun Chung would like to thank Simon Huttegger and Sean Ingham for serving as commentators for the paper at the Invited Paper session at the 2020 American Philosophical Association Pacific Division Meeting as well as the participants of the "Formal Approaches to Democratic Deliberation and Voting" session at the 2019 Annual Meeting of the PPE Society for their helpful feedback.

Received: March 20, 2018; revised: February 18, 2019; accepted: September 27, 2019.

[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, Nondictatorship, and Social Rationality.

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, 239–46).

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). According to deliberative democratic theorists, 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 1997, xix). This way of explaining the value of public deliberation and its connection to political justification presumes that there exist some procedure-independent criteria of rightness or correctness that the procedure of public deliberation is able to track. Many epistemic democrats (Estlund 1997; Landemore 2013; List and Goodin 2001; see also; Hong and Page 2004) hold this view.

Another group of scholars have claimed 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 (Christiano 1997, 252–3; Cohen 1997a, 76–7; Knight and Johnson 1997, 280; Rawls 1997, 133–4).

Finally, a number of scholars have argued that reasoned public deliberation may also complement (or even nullify the need for) aggregative voting mechanisms: by generating unanimous agreement (Elster 1997, 11–2; Habermas 1990); by "induc[ing] a shared understanding regarding the dimensions of conflict" (Knight and Johnson 1994, 282); or by inducing "single-peaked preferences" among the voters, which prevents majority rule from generating majority cycles (Dryzek and List 2003; see also; List et al. 2013).

In sum, there is an abundance of work in deliberative democratic theory that might be used 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 construct a formal theory of democratic deliberation itself. As some have pointed out, 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). The literature is beginning to fill this gap in the theory of deliberative democracy, as there is an increasing number scholars who are offering formal theories of democratic deliberation (Dietrich and List 2013; Hafer and Landa 2007; Landa and Meirowitz 2009; Patty 2008; Patty and Penn 2011, 2014; Perote-Pena and Piggens 2015). This article presents a formal theory of democratic deliberation that contributes to this growing line of research in a way summarized in the next section.

## OVERVIEW OF OUR THEORY

Here, we give an overview of our formal theory of deliberation. 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 opposing parties, each of whom seeks to employ rhetorical tactics to reach her favored position. 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. 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 position over another. The effectiveness of

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

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. A contribution of our approach is that it makes explicit the order in which positions are introduced and arguments are applied, or "protocol," so that we can impose different consistency conditions governing deliberative dynamics. Having isolated this aspect of deliberation, we examine the outcomes of different dynamics and see whether they satisfy the *desiderata* proposed by deliberative democratic theorists. We investigate three forms of deliberation that are distinguished by different deliberative dynamics.

First, we consider a *myopic discussion,* in which positions are introduced and arguments applied according to an exogenous protocol. An initial status quo position is given, and this evolves in a context-free way: if the position-argument pair given by the protocol is such that the position is superior to the status quo with respect to the argument, then it becomes the new status quo. It turns out that myopic discussion is 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 achieve many ideals of deliberative democracy.

Next, we provide a model of *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. 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 constructive discussion does better than 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 constructive discussion is to an extent 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 perfect 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 arguments 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. 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 preferences as well as the added strategic incentives among the participants that enable debate to meet the many lofty ideals of deliberative democracy.

## 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 arguments or reasons that can be given, 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.[3] In what follows, 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 $(x, y) \in P$, meaning that $x$ and $y$ are "in the relation" $P$. A binary relation $P$ on $X$ is *asymmetric* if for all positions $x, y \in X$, it is not the case that both $xPy$ and $yPx$ hold, so that they do not bear the relation to each other; it is *transitive* if for all positions $x$, $y$, $z \in X$, $xPy$ and $yPz$ together imply $xPz$; and it is *total* if for all distinct positions $x$, $y \in X$, either $xPy$ or $yPx$. Following the

[3] We take the set of positions as exogenously given, but a different line of inquiry would investigate the origins of the set of positions. It is possible that individual incentives to develop costly positions determine interesting or useful properties that could be leveraged in the analysis of deliberation.

standard convention, 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; such a relation can be represented by a ranking of positions, with no two positions tied at the same level. A position $x$ is *maximal* with respect to an asymmetric relation $P$ 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 $p: X \times X \rightarrow 2^A$ that associates a set $p(x, y) \subseteq A$ of reasons/arguments to each ordered pair $(x, y) \in X \times X$ of positions; here, we interpret $p(x, y)$ as the set of arguments that can be effectively used to support the position $x$ over position $y$. We can obtain any binary relation $P$ on $X$ as a special case by specifying that $A$ is a singleton set, say $A = \{1\}$, and then defining the mapping $p$ such that $p(x, y) = \{1\}$ holds if $xPy$, and $p(x, y) = \emptyset$ otherwise; thus, set-valued relations generalize the usual concept of a binary relation. We maintain the assumption that $p$ is *asymmetric*, in the sense that for all positions $x, y \in X$, we have $p(x, y) \cap p(y, x) = \emptyset$; in words, no argument that is effective for $x$ against $y$ is effective for $y$ against $x$.[4] In particular, for all positions $x \in X$, we have $p(x, x) = \emptyset$. That is, no argument 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 $a \in p(y, x)$, or equivalently, for all distinct $x, y \in X$, we have $p(x, y) \cup p(y, x) = A$. This means that an argument will always cut one way or the other between two positions, and it captures the idea that arguments can reflect fine distinctions between positions. For example, if an argument $a$ compares positions by some quantitive measure, it is enough that the measure is sufficiently fine grained that no two distinct positions are exactly at the same level. This condition may be considered restrictive, but this is a matter of interpretation: instead of an argument $a$ that is "incomplete," in the sense that it fails to compare two positions, we can often substitute a more refined argument that combines $a$ lexicographically with other criteria that can be used when $a$ does not apply.[5] We say $p$ is *transitive* if for all positions $x, y, z \in X$, 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 *constituent* relation $P^a$ on $X$ as follows: $xP^ay$ holds if and only if $a \in p(x, y)$. Clearly, as $p$ is asymmetric, each relation $P^a$ is asymmetric. Next, we state a Lemma showing 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. Thus, we can represent an asymmetric, set-valued relation $p$ by a collection $\{P^a | a \in A\}$ of asymmetric relations. The proof of the Lemma, along with all other formal results, is relegated to the Appendix: Technical Material.

**Lemma:** *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.*

Our model of arguments can be both related to and distinguished from other notable approaches that have been proposed in the literature. Patty and Penn (2011, 2014) take as primitives a finite set $X$ of alternatives, and a set $\pi$ of binary relations on $X$, where each $P$ belonging to $\pi$ is a possible "principle" that judges the merits of alternatives in $X$. Thus, their primitives are similar to ours, but their focus is different: rather than investigating the implications of different deliberative dynamics, they impose axioms on "procedures," which for each $(X, \pi)$ pair, determine a finite sequence of alternatives and a single principle $P \in \pi$ that is used to justify the final alternative in the sequence. We return to the work of these authors following Theorem 4, on constructive discussion.

Dung (1995) also offers a model of arguments, but his terminology is different than ours: his arguments are our positions, and he begins with a binary relation on arguments called "attacks." Given this relation, he proposes the concept of "preferred extension," which produces a set of arguments that is internally consistent and can be defended from outside attack. In fact, his conflict-free condition is the familiar concept of internal stability; and then his requirement that a preferred extension be a maximal admissible set adds a weakened form of external stability. Thus, the preferred extensions are closely related to the stable solutions (von Neumann and Morgenstern 1944): every stable solution is a preferred extension, but the latter exist even when the former do not.[6] Compared to our approach, Dung's analysis is based on a single relation on positions, rather than a collection of constituent relations for each argument, so (like the social choice concepts discussed at the end of the section) he does not exploit the full structure of a set-valued relation $p$; moreover, he imposes versions of internal and external stability, whereas we examine the implications of particular deliberative dynamics.

[4] We view asymmetry as unrestrictive. Consider an example in which $x$ is a tax decrease, and $y$ is a tax increase. One might be concerned that if $a$ is the economic growth argument, then one person might argue that $x$ is superior to $y$ on the basis of $a$, while another might argue that $y$ is superior to $x$ on the same grounds, apparently violating asymmetry. Here, we would say that there are really two arguments: $a'$ is that personal incomes and profits are higher, so that $a' \in p(x, y)$; and $a''$ is that environmental and social costs are lower, so that $a'' \in p(y, x)$.

[5] In the Appendix: Technical Material, we provide an example of this in the context of buying a home.

[6] In some cases, the preferred extension is empty—this violates the usual external stability condition, and is difficult to interpret in the context of deliberation.

Recall that an asymmetric relation is a linear order if it is total and transitive. Combining the observations of the Lemma, 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, the Lemma implies that we can summarize the effectiveness of the arguments by three linear orders, $P^f$, $P^c$, and $P^p$. For concreteness, we borrow from the classical Condorcet paradox,[7] and we assume 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 most fuel efficient (followed by the minivan and the sports car), the minivan is cheapest (followed by the sports car and luxury sedan), and the sports car has the best performance (followed by the luxury sedan and minivan). ||

Let us say that a position $x$ is *unassailable* if there is no position that is superior to it by 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.

To address the situation in which $UA$ is empty, given two positions, $x$ and $y$, we say $x$ *dominates* $y$, and write $x\bar{P}y$, if both of the following hold: $p(x, y) \neq \emptyset$, and for all $z \in X$, we have $p(z, x) \subseteq p(z, y)$.[8] That is, there is an argument for $x$ over $y$, and for all positions $z$, 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 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, $UD = \{x \in X | \nexists y \in X \text{ such that } y\bar{P}x\}$. Note that if a position is unassailable, then it is undominated. Hence, $UA \subseteq UD$. Theorem A.1, in the Appendix: Technical Material, shows that the dominance relation $\bar{P}$ is a partial order, implying that when $X$ is finite, an undominated position always exists, even if there is no unassailable one. Furthermore, 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.

Using the Lemma, when $X$ is finite and $p$ is total and transitive, each linear order $P^a$ has a position, denoted $x^a$, uniquely ranked at the top of the ordering. Such a position is clearly undominated, and these positions will possess the following strong stability property: for all $a \in A$ and all $y \in X \backslash \{x^a\}$, we have $a \in p(x^a, y)$. Informally, for every position $y$ distinct from $x^a$, the position $x^a$ is superior to $y$ by argument $a$. 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 minivan against both the luxury sedan and the sports car.

To delve more deeply into the structure of set-valued relations, and to facilitate the study of myopic discussions in the next section, we define the *projection* of $p$ as 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 $y$.[9] In terms of this relation, we can equivalently define the set $UA$ of unassailable positions as the set of maximal elements of $P^*$. The *transitive closure* of $P^*$ is denoted $P^\infty$ and defined as follows: for all positions $x, y \in X$, $xP^\infty y$ holds if and only if there is a path from $x$ to $y$, i.e., there exist a natural number $k$ and positions $x_1, \ldots, x_k \in X$ such that $xP^*x_1P^* \cdots x_{k-1}P^*x_k = y$. The transitive closure relation is transitive, as the name suggests, but it is not necessarily asymmetric. We say $x$ is *maximal* with respect to $P^\infty$ 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 $P^\infty$. Informally, if a position $x$ belongs to the top cycle, then whenever somebody constructs a chain of arguments that shows that some other position $y$ is superior to $x$, then it is always possible to counter this move by supplying another chain of arguments that shows that $x$ is superior to $y$.

[7] To be clear, while rankings in the Condorcet paradox represent voter preferences, they do not correspond to voter preferences—even by analogy—in our framework; see the discussion at the end of the section for connections to social choice theory.

[8] Note that since $p$ is irreflexive, a necessary condition for $x$ to dominate $y$ is that $p(y, x) = \emptyset$, i.e., $y$ is not superior to $x$ by any arguments; this condition is not sufficient, however.

[9] Equivalently, one can define $P^*$ as the union of the relations $P^a$.

As positions in the top cycle have this minimal degree of plausibility, 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. 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, but it can be very large, which may be problematic in the context of democratic deliberation when we have to make a specific policy choice.

**Discussion (Connections to Social Choice):** Several of the concepts defined above appear in a specialized form in the literature on social choice theory, specifically the tournament literature. Given a finite set $X$, a *tournament* is a binary relation $P$ on $X$ that is asymmetric and total. As discussed above, any tournament $P$ can be formalized as a set-valued relation, in which case the core, uncovered set, and top cycle of the tournament (Moulin 1986) coincide with the sets $UA$, $UD$, and $TC$ defined above; thus, our concepts specialize to familiar ideas in this setting. If the relation $P$ is asymmetric, but not necessarily total, or a *generalized tournament,* then again $UA$ corresponds to the core of a tournament, but it is known that there are multiple ways to extend the notions of the uncovered set and top cycle (Duggan 2013; Schwartz 1986). In fact, applying our definition of $UD$, we obtain the uncovered set of Gillies (1959), and our definition of $TC$ yields the GOCHA set of Schwartz (1986).

However, the structure imposed in our theory is richer than that of the tournament framework, where a single binary relation on $X$ is assumed. In our framework, we essentially begin with a collection of constituent relations, $P^a$. Recall, however, that our top cycle is determined by the projection $P^*$—which is just a binary relation on $X$. That is, $TC$ does not depend on the "internal structure" of $P^*$, making it similar to the top cycle from social choice theory. Note, however, that the projection $P^*$ may be asymmetric and total, but it need not satisfy either condition. Thus, our formulation of $TC$ extends the GOCHA set of Schwartz to relations violating asymmetry.[10] In contrast, our definition of the undominated set $UD$ does rely on the internal structure of $P^*$, in the sense that two set-valued relations, say $p$ and $p'$, may determine distinct undominated sets, even if they have the same projections. As we will see, our formulations of discussion (myopic and constructive) and debate in the remainder of the article all rely on this internal structure as well.

In fact, there is an implicit structure underlying the tournament framework not dissimilar to ours. In that literature, the tournament relation $P$ is often interpreted as being generated by voting in the following way: there is a number $n$ of voters and a profile $(P_1, \ldots, P_n)$ of voter preferences, where each $P_i$ is a linear order of $X$; then the relation $xPy$ is defined to hold if we have $xP_iy$ for at least $k$ voters, where $k$ is a fixed quota. It is assumed that $k > \frac{n}{2}$ to ensure asymmetry of $P$, and often it is assumed that $k = \frac{n+1}{2}$, so that voting is by majority rule, and no ties are possible, i.e., $P$ is total. In other words, the relation $P$ has internal structure generated by voting with respect to the quota $k$. In our framework, if we index arguments as $a_1, \ldots, a_n$, we have a profile $(P^{a_1}, \ldots, P^{a_n})$ of relations, one for each argument, and $xP^*y$ is defined to hold if we have $xP^{a_i}y$ for at least one argument.[11] If we view each argument as a voter, then our framework appears similar to the social choice one, with two differences. First, we in general allow $P^{a_i}$ to be any asymmetric relation; translated to the social choice context, this means we allow for voters with intransitive preferences. Second, because $xP^*y$ holds if $x$ is superior to $y$ for at least one argument, we essentially set the quota to $k = 1$ to generate $P^*$, consistent with our observation that $P^*$ may violate asymmetry. Thus, the internal structure of $P^*$ differs from that considered in the tournament literature, and our analysis of deliberation examines three different deliberative dynamics that are built upon this structure. ||

## 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 democratic theorists have suggested that deliberative democracy should, at least at the theoretical level, target unanimous agreement as its ultimate aim (Cohen 1997a; Elster 1997; Gaus 1997; Harbermas 1990). However, many critics have argued that full consensus or unanimous agreement is unlikely to be achieved in realistic circumstances, especially, in modern pluralistic societies (Knight and Johnson 2007; Rawls [1999] 2005; Elster 1997; Christiano 1997).

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 democratic theorists. tend to advert back to aggregative mechanisms in this situation. 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 democratic theorists not as something desirable in itself, but as a "necessary evil" that must be resorted to for practical purposes.

[10] In fact, our concept can be found by going back further to the literature on finite Markov chains (Doob 1953). Think of $X$ as a set of states, and assume that from each state $x$, we transition with equal probability over states $y$ such that $yP^*x$. This defines a Markov chain, and our TC is precisely the union of ergodic classes of this chain.

[11] We thank an anonymous referee for suggesting this interpretatoin.

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. One of the major characteristics of a deliberative democracy is that its process is *dynamic* in the sense that "[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:

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

If $x^m = z^{m+1}$, then we say $x^m$ 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 position $y$ 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 the 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:

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

In such a discussion, a position replaces the status quo by any argument according to 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. Intuitively, 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 | ⋯ |
| $a^m$ | f | f | c | c | p | p | ⋯ |
| $z^{m+1}$ | V | L | S | V | L | S | ⋯ |

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 two rounds in the above segment use the pairs $(L, p)$ and $(S, p)$. 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 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 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 $a$. When $p$ is total, however, the implications are sharper: every myopic discussion eventually reaches the top cycle and remains thereafter.

**Theorem 1 (Long Run Outcomes of Myopic Discussion):** *Assume the set $X$ of positions is finite. For every myopic discussion $\mathfrak{D}$ and every position $x \in X \backslash \Lambda(\mathfrak{D})$ outside the limit set and every argument $a \in A$, there is a position $y \in \Lambda(\mathfrak{D})$ in the limit set such that $a \notin p(x, y)$. Moreover, if the set-valued relation $p$ is total, then for every myopic discussion $\mathfrak{D}$, the limit set is contained in the top cycle: $\Lambda(\mathfrak{D}) \subseteq TC$.*

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 it 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 discourse essentially ends once the status quo has reached its final position. Thus, when a discussion is conclusive, we can at least say it meets one desideratum 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, there may be no unassailable position. This negative result is an immediate corollary of Theorem 1: 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 1 implies that for every other position $x$ and every argument $a$, we must have $a \notin p(x, y)$, implying that $y$ is unassailable. If there is no unassailable position, then myopic discussion cannot be conclusive and will cycle through the limit set *ad infinitum.*

**Corollary:** *Assume the set $X$ of positions is finite. If a myopic discussion $\mathfrak{D} = \{(x^m,\ a^m,\ z^m)\}_{m=1}^{\infty}$ is conclusive, then it concludes with some unassailable position $x$: there is a 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 endlessly, and it can be conclusive only under special circumstances. But how bad can it get? When $p$ is total, Theorem 1 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 thereafter. In fact, we use a protocol that is *rotating,* in the sense that for some natural number $n$, the first $n$ position-argument pairs are $(x^1, a^1)$, …, $(x^n, a^n)$, and the protocol repeats this pattern endlessly.[12]

**Theorem 2 (Indeterminacy of Myopic Discussion):** *Assume the set $X$ of positions is finite, and the set-valued relation $p$ 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: $\Lambda(\mathfrak{D}) = TC$.*

Theorem 2 implies that the limit set of a myopic discussion can coincide with the top cycle and, thus, be quite large. Worse yet, the top cycle can contain positions that are dominated, so an implication is the possibility that myopic discussion can visit dominated positions an infinite number of times. We provide an example illustrating this possibility in the Appendix: Technical Material following the proof of Theorem 2.

We end this section by noting an apparent analogy between Theorem 1 and 2 and results of Ordeshook and Schwartz (1987), who analyze the possible outcomes of sincere and strategic voting in binary voting agendas. The frameworks are conceptually distinct, as those authors work with voting agendas, in which the alternatives under consideration can depend on previous votes in an arbitrary way (see their Figures 3–5 for examples of the possible variety); moreover, they assume majority voting is represented by a tournament (an asymmetric, total relation), whereas we assume a set-valued relation $p$ and rely on the internal structure of $p$. The strongest parallel—albeit a superficial one—is provided by their Theorems 3 and 4, which prove that when voting is strategic, the set of outcomes obtainable by different voting agendas is contained in the top cycle, and that in fact all alternatives in the top cycle are obtainable by a specific subset of agendas. However, these results concern *strategic* voting, whereas myopic discussion is defined by a dynamic that is closest to *sincere* voting. For the case of sincere voting, Ordeshook and Schwartz show that almost every alternative can be obtained by using different agendas, including those outside the top cycle.[13]

## 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

[12] Formally, there exist $n$ positions $x_1, \ldots, x_n$ and arguments $a_1, \ldots, a_n$ such that for all $m$, $(x^m, a^m) = (x_{m \,(\mathrm{mod}\ n)}, a_{m\,(\mathrm{mod}\ n)})$.

[13] They show that any alternative that is not a Condorcet loser can be obtained—that set of alternatives comprising the "kitchen sink" set, in the terminology of Ordeshook and Schwartz.

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 large, but the limit set of myopic discussion can even 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 democratic deliberation is to deliver 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., $yP^ax$. That is, starting in round $m$, the new status quo for round $m + 1$ is

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

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: we prove that constructive discussion is conclusive, and in fact, it must conclude with a position that is maximal for some argument.[14] 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 3 (Conclusiveness of Constructive Discussion):** *Assume that the set X of positions is finite, and that the set-valued relation 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 is not an argument for y over x, i.e.,* $a \notin p(y, x)$. *Finally, if p is also total, then* $x = x^a$, *and x is undominated.*

Theorem 3 shows that in a constructive discussion, which allows deliberation to continue indefinitely, 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. Moreover, 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 concludes with agreement on some position; the final position to which a constructive discussion converges is maximally justified and, hence, best in terms of at least one argument; and in case $p$ is total, a constructive discussion will never conclude with a position that is dominated by another position. Nevertheless, Theorem 3 ensures that constructive discussion 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 $\Lambda = \{\lambda(\mathfrak{D}) | \mathfrak{D} \text{ is a constructive discussion}\}$ as the set of possible conclusions of constructive discussion, and we say constructive discussion is *path dependent* if $|\Lambda| > 1$; and otherwise it is *path independent.*

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. Combined with Theorem 3, an implication is that the positions that can be concluded from constructive discussion are exactly the maximal positions.

**Theorem 4 (Path Dependence of Constructive Discussion):** *Assume that the set X of positions is finite, and that the set-valued relation p is total and transitive. For every argument* $a \in A$, *the maximal position* $x^a$ *is reached as the conclusion of some constructive discussion, i.e., there exists a constructive discussion* $\mathfrak{D}$ *such that* $\lambda(\mathfrak{D}) = x^a$, *and thus the set of outcomes that can be concluded from constructive discussion is just the set of maximal positions*: $\Lambda = \{x^a | a \in A\}$. *In particular, if there exist arguments a, a'* $\in A$ *such that* $x^a \neq x^{a'}$, *then constructive discussion is path dependent.*

Next, we illustrate Theorem 4 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*.

**Example (Path Dependence of Car Purchases):** In the car purchase example, one possible constructive

[14] A close reading of the proof of Theorem 3 shows that the conclusiveness of constructive discussion does not rely on the assumption that constructive discussion is open.

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, as all arguments have been used to maximal effect. Thus, if the first three rounds of constructive discussion proceed as

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

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

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

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

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

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 dictated by Theorem 4. ∥

Theorem 4 might seem similar to Patty and Penn (2011)'s Theorem 11, in that each characterizes the outcomes determined by sequences of positions (or alternatives) satisfying different consistency conditions. As noted, Patty and Penn consider *procedures*, which map each pair $(X, \pi)$, where $\pi$ is a set of possible principles, to a decision sequence $(x^1, \ldots, x^M)$ in $X$ and a principle $P \in \pi$. They then impose two axioms on procedures: (i) no alternative $x$ earlier in the sequence is superior to any alternative $y$ later in the sequence according to $P$, i.e., we cannot have $xPy$ (internal stability); and (ii) for every alternative $z$ not in the sequence, there is some alternative $w$ in the sequence that is superior according to $P$, i.e., $wPz$ (external stability). Then, the final outcome $x^M$ is interpreted as being justified by the decision sequence. Their Theorem 11 shows that if we fix a single possible principle $\pi = \{P\}$, then the set of outcomes that can be justified by a legitimate procedure is the Banks set of $P$ (Banks 1985). Their analysis does not apply to constructive discussion: because a procedure determines a single principle that justifies the decision sequence, whereas constructive discussion involves multiple arguments, it is not possible to formalize constructive discussion as a procedure in the sense of Patty and Penn. Moreover, although constructive discussion is conclusive, and thus determines a finite sequence $(x^1, \ldots, x^M)$ of positions, these sequences do not satisfy anything like the internal stability axiom used by those authors.[15] Thus, our Theorem 4 and their Theorem 11 are quite different, both technically and in spirit.

As well, the path dependence result of Theorem 4 differs from those of McKelvey and Schofield, as they consider a very different mechanism: their indeterminacy is obtained by varying a binary voting agenda, with a multidimensional space of alternatives and a profile of preferences over alternatives for each voter as backdrop; indeed, for specific configurations of voter preferences in which the majority core is nonempty, their conclusion of indeterminacy does not hold. In contrast, Theorem 4 is formulated in a model of deliberation in which preferences do not play a role; rather, the indeterminacy arises from the effectiveness of arguments as the protocol of position-argument pairs is varied. And while McKelvey and Schofield show that for typical profiles, almost any alternative is possible, our indeterminacy conclusion is limited to the maximal positions.

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: with the same profile of individual preferences, we may reach radically different outcomes depending on the specific voting procedure adopted. 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 difficult for aggregative voting mechanisms to fully ground democratic justification or legitimacy, then Theorem 4 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 4 shows that Knight and Johnson's suspicion 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.

## A STRATEGIC MODEL OF DEBATE

One advantage of constructive discussion over myopic discussion, as Theorem 3 shows, is that it eventually reaches unanimous agreement on a single position. However, a critic might claim that, once we allow deliberation to proceed endogenously among parties who have diametrically opposed preferences over which position should be chosen, convergence or agreement

[15] In addition, constructive discussion determines the sequence $(x^1, \ldots, x^M)$ of positions only after a protocol is specified; so to formalize constructive discussion in their framework, we would have to associate with each $(X, \pi)$ pair a protocol used to generate the decision sequence.

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 define a debate as an equilibrium path of play in a particular extensive form game of perfect information that we call the *debate game.*[16] By perfect information, the participants of the debate game know all past moves taken 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 idealized circumstances by ideally rational agents (Cohen 1997a; Estlund 1997; Habermas 1990; Rawls 1997), and it is natural to assume, consistent with the assumption of perfect information, that ideally rational agents would remember what positions have been inserted or retained by what arguments in previous rounds of the debate game.

The players of the debate game are two participants, 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 game, actions $(x^1, a^1)$, …, $(x^m, a^m)$ determine a sequence $(z^2, \ldots, z^{m+1})$ of status quo positions as follows: for each $t = 1, \ldots, m$,

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

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. We can see that 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 for the first $m$ rounds,[17] 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 $m = 1, 2, \ldots$ As the debate game uses the same transition rule for the status quo as constructive discussion, Theorem 3 implies that every history is conclusive, in the sense that the status quo eventually settles on a single position that is repeated infinitely thereafter; formally, there is a position $x$ and round $m$ such that for all $k \geq m$, we have $z^k = x$.[18]

A (pure) *strategy* for player $i = 1, 2$ is a mapping, denoted $\sigma_i$, 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 $(\sigma_1, \sigma_2)$ then determines a *path of play*, which consists of the sequence of positions, arguments, and status quo positions generated by the choices of the players. Formally, this is the unique infinite history $h^\infty = \{(x^m, a^m, z^m)\}_{m=1}^\infty$ such that for each round $m$, (i) if $m$ is even, then we have $\sigma_1((x^1, a^1, z^1), \ldots, (x^m, a^m, z^m)) = (x^{m+1}, a^{m+1})$, 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 by assigning utilities to the *position* associated with that history. Specifically, since the status quo along any given history eventually settles down on a single position, we simply associate a history with the concluding position, and we specify that each participant receives utility from it. Formally, let $u_1(x)$ and $u_2(x)$ denote the utilities of participants 1 and 2, respectively, from any position $x$. We then assign a payoff to each infinite history $h^\infty$ as follows: we have noted that any given history $h^\infty$ concludes with a final position, say $x$, that eventually becomes the status quo and remains so; then we specify that participant 1's payoff from $h^\infty$ is $u_1(x)$, and participant 2's payoff from the history is $u_2(x)$. Thus, each player cares only about the conclusion of the debate game, and they evaluate the concluding position according to the utility functions $u_1$ and $u_2$.

Because the set $X$ of positions is finite, we can without loss of generality index it in the order of participant 1's preference, so that $X = \{x_1, \ldots, x_n\}$ and $u_1(x_1) < u_1(x_2) < \cdots < u_1(x_n)$. We assume that the participants of the debate game have opposing preferences, so the debate is competitive as well as constructive. Formally, 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. Given the above indexing, this means that $u_2(x_1) > u_2(x_2) > \cdots > u_2(x_n)$. 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 the move. Moreover, since we consider only pure strategies, we can without loss of generality (by a monotonic transformation of payoffs) further assume the game is zero sum, i.e., for all $x \in X$, we have $u_1(x) + u_2(x) = 0$.[19]

**Discussion (Competitive Debate):** An assumption that is almost universally endorsed by deliberative democratic theorists is that, in modern pluralistic

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

[17] 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.

[18] As remarked in footnote 13, the proof of Theorem 3 does not use the assumption that constructive discussion is open, and thus it applies equally well to the debate game.

[19] Nothing of substance relies on this monotonic transformation; we use it only to introduce the more familiar "zero-sum" terminology, and to more readily invoke results from zero-sum games.

societies, political deliberation occurs under conditions of irreconcilable moral and political disagreement. “A deliberative democracy is a pluralistic association” explains Cohen (1997a, 72), 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 explain 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 go further to claim that the whole purpose of democratic deliberation is to resolve “political conflict” (Knight and Johnson 1994, 285).

We seek a model of debate that reflects these conditions of irreconcilable moral and political disagreements, and the formal expression of the conditions is in the assumption of competitive payoffs. In our setting, this is equivalent to the assumption of zero-sum utility, i.e., $u_1(x) + u_2(x) = 0$ for all positions; again, because the payoffs of the two participants of the debate game are ordinal, any competitive game can be viewed as a zero-sum game, by a monotonic transformation that does not affect the set of equilibria.[20] We emphasize that the assumption of zero-sum utility for positions does not imply that the final position reached through the process of debate will 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 would run counter to the general spirit of deliberation sought by deliberative democratic theorists who see democratic deliberation as a device for achieving the common good (Cohen 1997a, 1997b; Elster 1997; Gutman and Thomson 2004). What is essential is the existence of *disagreement* in the model of debate, and thus we model debate as both constructive and competitive, in accordance with the deliberative environment assumed by deliberative democratic theorists. ‖

We analyze the debate game as a two-player, zero-sum game game of perfect 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 $\sigma_1$ and $\sigma_2$ 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 $\sigma_i^{'}$. 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 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, denoted $\lambda(\mathfrak{D})$.

Since constructive discussion, in general, is path dependent, we are now interested in whether 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 $\Lambda^* = \{\lambda(\mathfrak{D}) | \mathfrak{D} \text{ is a debate}\}$ 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

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

and

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

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 $x_\ell$ 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

[20] Observe that in a version of the debate game with general preferences over positions, the participants’ preferences over the *Pareto frontier* are nevertheless competitive: if two positions are Pareto efficient, then the participants must have opposing preferences over them. Thus, one intuition for our zero-sum assumption is that debate is restricted to the Pareto efficient positions.

the initial status quo. An immediate implication is that debate, in contrast to constructive discussion, does not suffer from the problem of path dependence.[21,22]

**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: $\Lambda^* = \{x^*\}$.*

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 $m \geq 4$.

For an intuitive story, 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 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

[21] In the Appendix: Technical Material, 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.

[22] The cheap-talk literature on deliberation with private information has shown that strategic incentives can create path dependencies, as the outcome can depend on the voting rule or order of message transmission. Theorem 5 does not contradict these results, as the mechanism considered in our paper is very different: that work considers cheap talk prior to a vote between two alternatives (Austen-Smith and Feddersen 2006) or messages from reputation-concerned experts prior to a choice between two alternatives by a single decision maker (Ottaviani and Sorensen 2001); in contrast, we consider deliberation in which an arbitrary, finite set of positions is narrowed down to a single one by a process of argumentation.

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 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 Habermas 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.[23]

## 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 legitimate decisions remains an educated guess. It is in this capacity that we believe formal analysis can empower normative philosophical theorizing.

In this article, 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 competitive debate—are consistent with the general characterizations of ideal deliberative procedure laid forth by deliberative democratic theorists; their underlying structure manifests equality, freedom of expression, fair equal opportunity to speak, and reciprocity. 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, the 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 article, 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

Arrow, Kenneth. (1951) 1963. *Social Choice and Individual Values*, 2nd edition. New Haven and London: Yale University Press.

Austen-Smith, David, and Timothy J. Feddersen. 2006. "Deliberation, Preference Uncertainty, and Voting Rules." *American Political Science Review* 100 (2): 209–17.

Banks, Jeffrey. 1985. "Sophisticated Voting Outcomes and Agenda Control." *Social Choice and Welfare* 4: 295–306.

Bohman, James, and William Rehg. 1997. *Deliberative Democracy: Essays on Reason and Politics*. Cambridge, MA: MIT Press.

Christiano, Thomas. 1997. "The Significance of Public Deliberation." In *Deliberative Democracy: Essays On Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 243–78.

Cohen, Joshua. 1997a. "Deliberation and Democratic Legitimacy." In *Deliberative Democracy: Essays on Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 67–92.

Cohen, Joshua. 1997b. "Procedure and Substance in Deliberative Democracy." In *Deliberative Democracy: Essays on Reason and*

[23] 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.

*Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 407–38.

Dietrich, Franz, and Christian List. 2013. “A Reason-Based Theory of Rational Choice.” *Noûs* 47 (1): 104–34.

Doob, J. L. 1953. *Stochastic Processes*. New York: John Wiley & Sons.

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

Duggan, John. 2013. “Uncovered Sets.” *Social Choice and Welfare* 41: 489–535.

Duggan, John, and Thomas Schwartz. 2000. “Strategic Manipulability without Resoluteness or Shared Beliefs: Gibbard-Satterthwaite Generalized.” *Social Choice and Welfare* 17: 85–93.

Dung, Phan Minh. 1995. “On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and *N*-Person Games.” *Artificial Intelligence* 17 (2): 321–57.

Elster, Jon. 1997. “The Market and the Forum: Three Varieties of Political Theory.” In *Deliberative Democracy: Essays on Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 3–34.

Estlund, David. 1997. “Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority.” In *Deliberative Democracy: Essays on Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 173–204.

Gaus, Gerald. 1997. “Reason, Justification, and Consensus: Why Democracy Can’t Have It All.” In *Deliberative Democracy: Essays on Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 205–42.

Gibbard, Allan. 1973. “Manipulation of Voting Schemes: A General Result.” *Econometrica* 41 (4): 587–601

Gillies, Donald B. 1959. “Solutions to General Non-Zero-Sum Games.” In *Contributions to the Theory of Games IV: Annals of Mathematics Studies*. Vol. 40, eds. Albert William Tucker and Robert Duncan Luce. Princeton, NJ: Princeton University Press.

Gutman, Amy, and Dennis Thomson. 2004. *Why Deliberative Democracy?* Princeton, NJ: Princeton University Press.

Habermas, Jürgen. 1975. *Legitimation Crisis*, trans. Thomas McCarthy. Boston: Beacon Press; London: Heinemann.

Habermas, Jürgen. 1990. “Discourse Ethics: Notes on a Program of Philosophical Justification.” In *Jürgen Habermas, Moral Consciousness and Communicative Action*, trans. C. Lenhardt and S.W. Cicholsen. Cambridge, MA: MIT Press, 43–115.

Hafer, Catherine, and Dimitri Landa. 2007. “Deliberation as Self-Discovery and Institutions for Political Speech.” *Journal of Theoretical Politics* 19 (3): 329–60.

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): 16385–9.

Knight, Jack, and James Johnson. 1994. “Aggregation and Deliberation: On the Possibility of Democratic Legitimacy.” *Political Theory* 22 (2): 277–96.

Knight, Jack, and James Johnson. 1997. “What Sort of Equality Does Deliberative Democracy Require?” In *Deliberative Democracy: Essays on Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 279–320.

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.

Laffond, Gilbert, Jean-Francois Laslier, and Michel Le Breton. 1993. “The Bipartisan Set of a Tournament Game.” *Games and Economic Behavior* 5: 182–201.

Laffond, Gilbert, Jean-Francois Laslier, and Michel Le Breton. 1997. “A Theorem on Symmetric Two-Player Zero-Sum Games.” *Journal of Economic Theory* 72: 426–31.

Landa, Dimitri, and Adam Meirowitz. 2009. “Game Theory, Information, and Deliberative Democracy.” *American Journal of Political Science* 53 (2): 427–44.

Landemore, Helene. 2013. *Democratic Reason*. Princeton, NJ: Princeton University Press.

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

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.

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

May, Kenneth. 1952. “A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision.” *Econometrica* 20: 680–4.

McKelvey, Richard. 1976. “Intransitivities in Multi-Dimensional Voting Models and Some Implications for Agenda Control.” *Journal of Economic Theory* 12: 472–82.

McKelvey, Richard. 1979. “General Conditions for Global Intransitivites in Formal Voting Models.” *Econometrica* 47: 1085–1112.

Moulin, Hervé. 1986. “Choosing from a Tournament.” *Social Choice and Welfare* 3: 271–91.

Ordeshook, Peter, and Thomas Schwartz. 1987. “Agendas and the Control of Political Outcomes.” *American Political Science Review* 81: 179–200.

Osborne, Martin, and Ariel Rubinstein. 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press.

Ottaviani, Marco, and Peter Sorensen. 2001. “Information Aggregation in Debate: Who Should Speak First?” *Journal of Public Economics* 81 (3): 393–421.

Patty, John. 2008. “Arguments-Based Collective Choice.” *Journal of Theoretical Politics* 20 (4): 379–414.

Patty, John, and Elizabeth Maggie Penn. 2011. “A Social Choice Theory of Legitimacy.” *Social Choice and Welfare* 36: 365–82.

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

Perote-Peña, Juan, and Ashley Piggens. 2015. “A Model of Deliberative and Aggregative Democracy.” *Economics and Philosophy* 31: 93–121.

Plott, Charles. 1967. “A Notion of Equilibrium and Its Possibility under Majority Rule.” *The American Economic Review* 57: 787–806.

Rawls, John. (1993) 2005. *Political Liberalism*, Expanded edition. New York: Columbia University Press.

Rawls, John. 1997. “The Ideal of Public Reason.” In *Deliberative Democracy: Essays on Reason and Politics*, eds. James Bohman and William Rehg. Cambridge, MA: MIT Press, 93–141.

Riker, William. 1982. *Liberalism Against Populism*. Long Grove, IL: Waveland Press.

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.

Schofield, Norman. 1978. “Instability of Simple Dynamic Games.” *The Review of Economic Studies* 45: 575–94.

Schwartz, Thomas. 1986. *The Logic of Collective Choice*. New York, New York: Columbia University Press.

von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

## Appendix: Technical Material

This appendix contains technical material omitted from the body of the paper. We begin by proving the Lemma, which decomposes properties of a set-valued relation into properties of its constituent relations.

**Lemma:** *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 $xP^az$, 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 $xP^ayP^az$, and transitivity of $P^a$ implies $xP^az$, 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. □

Next, we present an example that illustrates the permissiveness of the assumption that a set-valued relation is total: namely, by refining arguments appropriately, it is often possible to transform a situation in which the condition is violated into one in which it is satisfied.

**Example (Buying a Home):** There are three alternatives under consideration for the purchase of a new home: two downtown apartments ($B$ and $C$) and one suburban house ($H$). Two possible criteria for choosing between these options are square footage ($f$) and speed of elevator ($e$). Let's assume that the house is bigger than apartment $B$, which is bigger than apartment $C$, and assume that the elevator to $C$ is faster than the elevator to $B$; of course, the house has no elevator, so it cannot be compared to the apartments by this criterion. Then the set of positions and set of arguments are

- $X = \{B, C, H\}$
- $A = \{e, f\}$.

and we have the following:

$$p(B,C) = \{f\}, p(B,H) = \emptyset, p(C,B) = \{e\}, p(C,H) = \emptyset,$$
$$p(H,B) = p(H,C) = \{f\}.$$

In particular, we have $e \notin p(B,H) \cup p(H,B)$, so the set-valued relation $p$ is not total. Here, $e$ is interpreted as "has a faster elevator than," but we can reformulate the example with a more refined argument $e'$, which means "has a faster elevator than, or if that doesn't apply, has greater square footage than." Formally, we model the set of arguments as $A'$ = $\{e', f\}$, and we reformulate $p$ as $p'$, defined as

$$p'(B,C) = \{f\}, p'(B,H) = \emptyset, p'(C,B) = \{e'\}, p'(C,H) = \emptyset,$$
$$p'(H,B) = \{e', f\}, p'(H,C) = \{e', f\}.$$

This reinterpretation is harmless, in the sense that it does not affect the top cycle or undominated set, but it allows us to apply our results for total, set-valued relations to the modified model. ||

Next, we note that the dominance relation $\bar{P}$ 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 A1:** *The dominance relation $\bar{P}$ is a partial order; and thus if the set $X$ of positions is finite, then there is an undominated position. Moreover, given any position $x \in X$, if $p(x,y) \neq \emptyset$ for every $y \in X \backslash \{x\}$, then $x$ is undominated. Assuming the set-valued relation $p$ is total, then for all positions $x$, $y \in X$, $x\bar{P}y$ holds if and only if for all arguments $a \in A$, we have $xP^ay$; and thus if $x$ is undominated, then for every other position $y \in X \backslash \{x\}$, we have $p(x,y) \neq \emptyset$.*

*Proof:* To see that $\bar{P}$ is asymmetric, suppose otherwise to deduce a contradiction. Then there exist positions $x, y \in X$ such that $x\bar{P}y$ and $y\bar{P}x$. By $y\bar{P}x$, it follows that for all $z \in X$, we have $p(z,y) \subseteq p(z,x)$. Setting $z = x$, and using irreflexivity of $p$, this implies $p(x,y) \subseteq p(x,x) = \emptyset$. But $x\bar{P}y$ implies $p(x,y) \neq \emptyset$, a contradiction. Thus, $\bar{P}$ is indeed asymmetric. For transitivity, consider any positions $x, y, z \in X$ and assume $x\bar{P}y\bar{P}z$. Then there exists $a \in p(x,y) \subseteq p(x,z)$ (now using $y\bar{P}z$ 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 $x\bar{P}z$. Thus, $\bar{P}$ is a partial order. If $X$ is finite, it follows immediately that $\bar{P}$ admits a maximal element, i.e., an undominated position. Now, consider any position $x \in X$, and assume that for all $y \in X \backslash \{x\}$, 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. □

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, $P^\infty$, 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 $(Y \backslash \{x\})P^\infty x$; and there is no superset of $Z \supsetneq Y$ such that for all $y \in Y$, we have $(Z \backslash \{y\})P^\infty y$. Roughly, a component is maximal (with respect to set inclusion) among sets that bear $P^\infty$ (i.e., the transitive closure of $P^*$) to each of their elements. For example, assume $X = \{x, y, z, w\}$ and $P^* = \{(x,y), (y,z), (z,x), (y, w)\}$. Here, the unique component is $Y = \{x, y, z\}$. To verify that $Y$ is a component, note that $yP^*zP^*x$, so $Y \backslash \{x\}P^\infty x$, and by symmetry, it follows that the first condition in the definition is fulfilled; and the only superset is $Z = X$, and since $w$ does not bear $P^*$ to $x$, $y$, or $z$, we do not have $ZP^\infty x$.

Next, we show that the top cycle consists of the union of components; moreover, if $X$ is finite and a position $x \in X \backslash TC$ 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 A2:** *The top cycle is the union of components, i.e., $TC$ $= \cup\{Y | Y$ is a component$\}$. Moreover, if the set $X$ of positions is* *finite, then for every position $x \in X\backslash TC$ outside the top cycle, there is a component $Y$ such that $YP^{\infty}x$. Finally, if the set-valued relation $p$ 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 $Y = \{y \in X|yP^{\infty}x\} \cup \{x\}$. If $Y = \{x\}$, then it is clearly a component, so assume $Y$ contains at least one position distinct from $x$. Note that for all $y \in Y\backslash\{x\}$, because $x$ is maximal with respect to $P^{\infty}$, we have $xP^{\infty}y$. Now consider $y \in Y$ and $z \in Y\backslash\{y\}$. 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 $(Y\backslash\{y\})P^{\infty}y$. Given $y \in Y$ and $z \in X\backslash Y$, 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 $(Y\backslash\{y\})P^{\infty}y$, which implies $xP^{\infty}y$. To show that $x$ belongs to the top cycle, it then suffices to rule out $y \in X\backslash Y$. Suppose toward a contradiction that $y \notin Y$. In case $Y = \{x\}$, we can set $Z = Y \cup \{y\}$ to arrive at $(Z\backslash\{x\})P^{\infty}x$, contradicting the assumption that $Y$ is a component. 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 $(Z\backslash\{z\})P^{\infty}z$, 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 $x \in X\backslash TC$. Let $Y = \{y \in X|yP^{\infty}x\}$, which is nonempty and finite. We claim that there is a maximal element of $P^{\infty}$ 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 $P^{\infty}$, 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 $Z = \{z \in X|zP^{\infty}y^*\} \cup \{y^*\}$. By the initial argument of the proof, it follows that $Z$ is a component, and by transitivity of $P^{\infty}$, 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 $yP^*z$. Consider any $x \in Z$ and any $y' \in (Y \cup Z)\backslash\{x\}$. In case $y' \in$ Z, then $y'$ $P^{\infty}x$ follows immediately from the fact that Z is a component, so assume $y' \in$ Y. In case $y' \neq y$ and $x = z$, then we have $y'P^{\infty}yP^{\infty}x$; and in case $y' \neq y$ and $x \neq z$, then we have $y'P^{\infty}yP^{\infty}zP^{\infty}x$. In all cases, we have $y'P^{\infty}x$. Since $y' \in Y$ and $x \in Z$ were arbitrary, it follows by definition of a component that $Y \cup Z \subseteq Z$, i.e., $Y \subseteq Z$. Since $Z$ is a component, we therefore have $zP^{\infty}y$, and a symmetric argument implies $Z \subseteq Y$, i.e., $Y = Z$. We conclude that there is a single component, and by the first part of the theorem, this coincides with the top cycle. We have already noted that $UA \subseteq UD$. Now, consider any undominated position $x \in UD$. By Theorem A1, it follows that for all $y \in X\backslash\{x\}$, we have $xP^*y$, and thus $xP^{\infty}y$, and this implies $x \in TC$. We conclude that $UD \subseteq TC$. □

Next, we prove Theorems 1–5 from the body of the paper.

**Theorem 1 (Long Run Outcomes of Myopic Discussion):** *Assume the set $X$ of positions is finite. For every myopic discussion $\mathfrak{D}$ and every position $x \notin X\backslash\Lambda(\mathfrak{D})$ outside the limit set and every argument $a \in A$, there is a position $y \in \Lambda(\mathfrak{D})$ in the limit set such that $a \notin p(x, y)$. Moreover, if the set-valued relation $p$ is total, then for every myopic discussion $\mathfrak{D}$, the limit set is contained in the top cycle: $\Lambda(\mathfrak{D}) \subseteq TC$.*

*Proof:* First, assume $X$ is finite, and let $\mathfrak{D}$ be a myopic discussion. Consider any position $x \in X\backslash\Lambda(\mathfrak{D})$ 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 $z \in X\backslash\Lambda(\mathfrak{D})$, 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 $m = \max\{m_z|z \in Z\backslash\Lambda(\mathfrak{D})\}$, 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 $y \in X\backslash TC$, 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 $z^{m+1} = x \in TC$. Then for all $n > m$, the status quo remains in the top cycle, which implies $\Lambda(\mathfrak{D}) \subseteq TC$. □

**Theorem 2 (Indeterminacy of Myopic Discussion):** *Assume that the set $X$ of positions is finite, and that the set-valued relation $p$ 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 $y \in X\backslash\{x\}$, 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 $x_1P^*x_2P^*\cdots x_nP^*x_1$. 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 $x_nP^*x_{n-1}P^*\cdots x_1P^*x_n$. Let $\mathcal{H}$ denote the collection of Hamiltonian sets. By Theorem A2, 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 $(TC\backslash\{y\})P^{\infty}y$ and $(TC\backslash\{x\})$ $P^{\infty}x$, which yields $xP^{\infty}yP^{\infty}x$. This gives us a $P^*$-cycle containing both $x$ and $y$, i.e., alternatives $x_1,\ldots, x_k \in X$ such that $x_1P^*x_2P^*\cdots x_kP^*x_1$ with $x = x_i$ and $y = x_j$ for some $i, j = 1,\ldots, k$. By choosing the shortest such cycle we ensure that the alternatives $x_1,\ldots, x_k$ are distinct, and then $Y = \{x_1,\ldots, x_k\}$ is Hamiltonian, and thus $\mathcal{H}$ is nonempty. Since it is finite, we can choose a Hamiltonian set $Y \in \mathcal{H}$ that is maximal among $\mathcal{H}$ with respect to set inclusion. We claim that $Y = TC$, for suppose toward a contradiction that $TC \setminus Y \neq \emptyset$. We discern two cases.

Case 1: for all $x \in TC\backslash Y$, 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 $x \in TC\backslash Y$ 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 $(TC\backslash\{y\})P^{\infty}y$, so there are positions $x \in TC \setminus Y$ and $y \in Y$ such that $x P^* y$, 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 $x_iP^*x$. Without loss of generality, assume $i = 1$, so $x_1P^*x$. By assumption, we have $xP^*x_2$, and thus we have $x_1P^*xP^*x_2\cdots P^*x_kP^*x_1$, but then $Y \cup \{x\}$ is Hamiltonian, contradicting maximality of $Y$. Case 1.2: there is no $i = 1,\ldots, k$ such that $x_iP^*x$. Since $(TC\setminus\{x\})P^\infty x$, there exist distinct $y_1, \ldots, y_\ell \in X$ with $y_1 \in Y$ and $y_1P^*y_2P^*\cdots y_\ell P^*x$. 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

$$x_1P^*\, x_2P^* \;\cdots x_i\, P^*\, y_{j+1}P^* \;\cdots y_\ell P^*\, xP^*\, x_{i+1}\, P^* \;\cdots\; x_2P^*x_1$$

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, $x_{i-1}P^*xP^*x_i$, and thus we have

$$x_1P^*x_2P^* \cdots x_{i-1}P^*xP^*x_iP^*x_{i+1} \cdots x_kP^*x_1,$$

but then $Y \cup \{x\}$ is Hamiltonian, contradicting maximality of $Y$. Case 2.2: $i < j$. Without loss of generality, assume $i = 1$, so $xP^*x_1$. Then we can choose $j$ to be as high as possible subject to the constraint that not $xP^*x_j$. Since $p$ is total, this implies $x_jP^*x$, and identifying $x_{k+1}$ with $x_1$, we can then write $x_jP^*xP^*x_{j+1}$. 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 $x_{n_1}P^*x_{n_1-1}P^* \cdots x_2P^*x_1P^*x_{n_1}$. 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 $xP^ay$ 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 $x^1P^*z^1$ implies $z^2 = x^1 \in TC$, or $z^2 = z^1$; and if $z^1 \notin TC$, then this holds because $x^1P^*z^1$, which again implies $z^2 = x^1 \in TC$. Then the status quo remains in the top cycle thereafter. Next, we claim that for every multiple of $n = n_1 + n_2$, say $\alpha n$ (where $\alpha$ 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 $\alpha n$ 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 $\alpha$ is an arbitrary positive integer, it follows that $x$ belongs to the limit set of $\mathfrak{D}$, as required. □

Next, we provide an example to demonstrate the possibility that myopic discussion can visit dominated positions an infinite number of times.

**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 $xP^ay$, 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 2, there is a rotating protocol such that every myopic discussion for that protocol cycles through the four positions endlessly. ||

**Theorem 3 (Conclusiveness of Constructive Discussion):** *Assume that the set $X$ of positions is finite, and that the set-valued relation $p$ 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 $y \in X$, $a$ is not an argument for $y$ over $x$, i.e., $a \notin p(y, x)$. Finally, if $p$ is also total, then $x = x^a$, and $x$ 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 $\left(x^{m_k-1}, a^{m_k-1}, z^{m_k-1}\right) = \left(x^{m_\ell-1}, a^{m_\ell-1}, z^{m_\ell-1}\right)$. 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.,

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

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

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

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 $y \in X \backslash \{x\}$, we have $a \in p(x, y)$, which is nonempty, and Theorem A1 implies that $x$ is undominated. □

**Theorem 4 (Path Dependence of Constructive Discussion):** *Assume that the set $X$ of positions is finite, and that the set-valued relation $p$ is total and transitive. For every argument $a \in A$, the maximal position $x^a$ is reached as the conclusion of some constructive discussion, i.e., there exists a constructive discussion $\mathfrak{D}$ such that $\lambda(\mathfrak{D}) = x^a$, and thus the set of outcomes that can be concluded from constructive discussion is just the set of maximal positions: $\Lambda = \{x^a | a \in A\}$. In particular, if there exist arguments $a$, $a' \in A$ 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., $\tilde{A} = \left\{\tilde{a} \in A | x^{\tilde{a}} = x^a\right\}$. If there is no argument $a'$ such that $x^{a'} \neq x^a$, then Theorem 3 immediately yields $\lambda(\mathfrak{D}) = x^a$ for every constructive discussion, so henceforth assume that $A \backslash \tilde{A} \neq \emptyset$. Letting $k = |\tilde{A}|$ and $\ell = |A|$, we can index $A \backslash \tilde{A}$ as $A \backslash \tilde{A} = \{a_1, \ldots, a_{\ell-k}\}$ 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 $(x^a, a_m)$ 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 3, 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., $m_{a'} = \min\{m | x^m = x^{a'}\}$. 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 $a' \in A \backslash \tilde{A}$ 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 $a' \in A \backslash \tilde{A}$, 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 $\Lambda = \{x^a | a \in A\}$, and thus if there are arguments with distinct top ranked positions, then constructive discussion is path dependent. □

**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: $\Lambda^* = \{x^*\}$.*

*Proof:* 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

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

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 $x^{m-1} \notin X^{a^{m-1}}\left(h^{m-1}, z^m\right)$, 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.,

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

and let $\alpha(h^{m-1}, z^m) = |A(h^{m-1}, z^m)|$ be the number of such arguments. Let $X^*(h^{m-1}, z^m) = \{x^a \in X | a \in A(h^{m-1}, z^m)\}$ 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 $x^{m-1} \neq z^{m-1}$, 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 $x^{m-1} = z^{m-1}$, we have already noted $x^{m-1} \notin X^{a^{m-1}}\left(h^{m-1}, z^m\right)$. 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, $X^{a^{m-1}}\left(h^{m-1}, z^m\right) = \emptyset$, 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

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,1}(x) \leq \frac{\alpha\left(h^{m-1}, z^m\right)}{2}, \tag{A1}$$

and

$$\sum_{a \in A(h^{m-1}, z^m)} I_{a,2}(x) \leq \frac{\alpha\left(h^{m-1}, z^m\right)}{2}. \tag{A2}$$

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 (A1) and (A2). We set $x^*(h^{m-1}, z^m)$ equal to the unique position $x \in X^*(h^{m-1}, z^m)$ satisfying (A1) and (A2), or, if there are two such positions, say $x_k$ and $x_\ell$ 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 unique subgame perfect equilibrium outcome is the compromise position $x^*(h^{m-1}, z^m) = x^a$. 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 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 subgame perfect equilibrium outcome, as claimed.

Next, assume the claim is true when the number of active arguments is 1, 2,..., $k$. Formally, assume that for all $(h^{m-1}, z^m)$ with $|A(h^{m-1}, z^m)| \leq k$, the unique subgame perfect 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 subgame perfect 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 $\underline{a} \in A\left(h^{m-1}, z^m\right)$ satisfy $\underline{x} = x^{\underline{a}}$, and suppose that player 1 justifies $\underline{x}$ by argument $\underline{a}$. Let $h^m = \left(h^{m-1}, \underline{x}, \underline{a}\right)$ 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

$$A\left(h^m, z^{m+1}\right) = A\left(h^{m-1}, z^m\right) \backslash \{\underline{a}\}.$$

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

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,1}(x^*) \leq \frac{\alpha\left(h^m, z^{m+1}\right)}{2} \tag{A3}$$

and

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) \leq \frac{\alpha\left(h^m, z^{m+1}\right)}{2}. \tag{A4}$$

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

$$\frac{\alpha\left(h^m, z^{m+1}\right)}{2} = \frac{\alpha\left(h^{m-1}, z^m\right)}{2} - \frac{1}{2}, \tag{A5}$$

it follows that $x^*$ satisfies (A1).

Moreover, $\alpha(h^m, z^{m+1})$ is even, and thus there may be one or two positions satisfying (A3) and (A4). If there is just one, then it must appear at the top of at least two arguments, and thus, the inequality in (A4) holds strictly at $x^*$, i.e.,

$$\sum_{a \in A(h^m, z^{m+1})} I_{a,2}(x^*) < \frac{\alpha\left(h^m, z^{m+1}\right)}{2}. \tag{A6}$$

If there are two positions satisfying (A3) and (A4), 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

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

where the second inequality follows from (A6), and the equality follows from (A5), and therefore $x^*$ satisfies (A2).

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 $\bar{x}$ minimize $u_2$ over $X^*(h^m, z^{m+1})$, and let $\bar{a}\in A\left(h^m, z^{m+1}\right)$ satisfy $\bar{x} = x^{\bar{a}}$. Then by a symmetric argument, player 2 can insert $\bar{x}$ by argument $\bar{a}$ and obtain the compromise position $x^*$. Since player 2's equilibrium strategy is optimal at $h^m$, it follows that $u_2(x')\ge u_2(x^*)$, and we deduce that $u_1(x')\le u_1(x^*)$. 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

$$A\left(\tilde{h}^m, \tilde{z}^{m+1}\right) = A\left(h^{m-1}, z^m\right)\backslash\{\tilde{a}\},$$

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

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

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

$$\sum_{a\in A\left(\tilde{h}^m,\tilde{z}^{m+1}\right)} I_{a,1}(\tilde{x}^*)\le\frac{\alpha\left(\tilde{h}^m,\tilde{z}^{m+1}\right)}{2}, \qquad \textbf{(A7)}$$

and

$$\sum_{a\in A\left(\tilde{h}^m,\tilde{z}^{m+1}\right)} I_{a,2}(\tilde{x}^*)\le\frac{\alpha\left(\tilde{h}^m,\tilde{z}^{m+1}\right)}{2}. \qquad \textbf{(A8)}$$

If $u_1(\tilde{x}) < u_1(x^*)$, then $x^*$ satisfies (A7) and (A8), and it follows that the compromise position at $\left(\tilde{h}^m, \tilde{z}^{m+1}\right)$ 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 (A3) and (A7) evaluated at $x^*$, we have

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

and comparing the left-hand sides of (A4) and (A8), we have

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

It follows that for every solution $x$ to (A7) and (A8), we have $u_1(x) \le u_1(x^*)$, and in particular, $u_1(x^*)\ge u_1(\tilde{x}^*)$. 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 (A3) and (A4), since $\alpha(h^m, z^{m+1})$ is odd. Since $u_1(x^*)\ge u_1(\underline{x})$, 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

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

which implies that $x^*$ satisfies (A1). Since $u_2(\underline{x})\ge u_2(x^*)$, 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

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

and thus $x^*$ satisfies (A2).

There may be one or two positions satisfying (A1) and (A2). 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 (A1) must hold with equality at $x^*$, i.e.,

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

But this implies

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

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 $\bar{x}$ by $\bar{a}$, 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. □

Finally, we end with a discussion of the Bipartisan set of Laffond, Laslier, and Le Breton (1993) and an extension of this concept to a model of debate between two win-motivated participants.

**Discussion (Bipartisan Debate):** Assuming a finite number of possible alternatives, recall that a tournament on $X$ is an asymmetric, total relation $P$. Laffond, Laslier, and Le Breton (1993) consider a two-player, symmetric, zero-sum game between two parties defined as follows: if party 1 chooses $x$ and party 2 chooses $y$, then party 1 receives a payoff of 1 if $xPy$, a payoff of $-1$ if $yPx$, and zero if $y = x$; and party 2's payoff is negative one times this. That is, interpreting $xPy$ as the situation in which alternative $x$ beats $y$, each party wants to beat the other. It is known that this game has a pure strategy equilibrium if and only if there is a Condorcet winner, i.e., an alternative $x$ such that for all $y \neq x, xPy$. In the absence of such a winning alternative, there is no pure strategy equilibrium, but the authors show that there is a unique mixed strategy equilibrium, and the support of the equilibrium strategy, i.e., the set of alternatives that are chosen with positive probability, is the *Bipartisan set* of the tournament.

An analogue to the bipartisan set can be defined in the setting of a debate between two participants. One simple way to do this is to let each the participants choose positions simultaneously; given position $x$ for participant 1 and position $y$ for participant 2, we can specify a payoff of 1 to participant 1 if there is an argument for $x$ over $y$ but not the reverse, i.e., $p(x,y) \neq \emptyset = p(y,x)$; and we specify that participant 1's payoff is $-1$ if $p(y,x) \neq \emptyset = p(x,y)$; and participant 2's payoff is negative one times this. Because $P^*$ is not a tournament, the result of Laffond, Laslier, and Le Breton (1993) does not apply, and this game can have multiple equilibria. Because the game is zero sum, however, there will be a unique mixed strategy equilibrium with a support set that contains the support of all other equilibria. This approach does not account for the possibility that $x$ may be stronger relative to $y$ than vice versa, in the sense that there are more arguments that favor $x$ than favor $y$.

Interestingly, we can extend the approach to capture this. Assume there are a finite, odd number of arguments, and that $p$ is total. For each pair of positions $x$ and $y$, let $\pi(x, y)$ be the number of arguments in favor of $x$ over $y$, minus the number in favor of $y$ over $x$, i.e.,

$$\pi(x,y) = \#p(x,y) - \#p(y,x).$$

By the preceding assumptions, $\pi(x,y)$ is odd whenever $x$ and $y$ are distinct, and then it follows from Laffond, Laslier, and Le Breton (1997) that there is a *unique* mixed strategy equilibrium of the game, and the support set of the equilibrium strategy provides a natural extension of their concept. Obviously, the support of this equilibrium strategy will be a singleton consisting of the compromise position $x^*$ only in the very special case that $\pi(x^*, y) > 0$ for all $y \neq x^*$; thus, the equilibrium incentives in this simultaneous-move game differ markedly from the debate game analyzed in Section 6. ||