Game Theory, Information, and Deliberative Democracy*

Dimitri Landa† Adam Meirowitz‡

September 25, 2006

Abstract

The game-theoretic analysis of deliberation and the normative theory of deliberative democracy share an interest in deliberation but have, by and large, been developing in mutual isolation. We explore the central methodological issues at the core of the relationship between these approaches and articulate an account of micro-foundations for the democratic theory based on their synthesis. We then confront the arguments raised by normative theorists against the relevance of the game-theoretic work. We contend that, with a suitably wide notion of rationality and a broad set of motivations, the game-theoretic tradition is particularly well-suited for generating insights about the feasibility of deliberative institutions and practices. The role of normative theories of deliberation is to provide justifications for behavioral and institutional desiderata and their optimal tradeoffs.

*We thank Chris Achen, Larry Bartels, Tim Feddersen, Sandy Gordon, Cathy Hafer, and David Stasavage for their helpful comments on the earlier drafts of this paper.

†Corresponding author. Assistant Professor of Politics, New York University. Address: Department of Politics, New York University, 726 Broadway, 7th floor, New York, New York 10003. E-mail: dimitri.landa@nyu.edu.

‡Associate Professor of Politics, Princeton University. Address: Department of Politics, Corwin Hall, Princeton University, Princeton, NJ 08544. E-mail: ameirowi@princeton.edu.

Introduction

Democracy is minimally defined as a form of governance in which policy decisions are made by a majority vote of the citizens. While useful as a rough way of classifying polities, this definition turns out, on closer examination, to be both ambiguous and radically incomplete. The main thrust of the critique of minimal democracy developed in contemporary democratic theory is that voting is not the best, or at least not the only, political mechanism determining whether policy decisions conform to the interests of the citizens. A key political mechanism that also serves that role and that is missing from the minimalist view of democracy is deliberation, and the appreciation of the effects of this mechanism is changing the way scholars of democracy think about democratic institutions.

In revealing correct, fuller, or simply better organized information, deliberation provides an opportunity for participants to arrive at more considered judgments of their own and to affect collective decision-making by influencing the judgments of others. Its consequences may affect what happens in a voting booth or in a legislative or a judicial chamber, or in the way we approach a personal moral conundrum. A political decision-making process that fails to create the opportunity for or to take advantage of these benefits of deliberation is bound to raise questions about the legitimacy of the resulting outcomes (Manin 1987; Habermas 1996; Cohen 1996).

Apart from the immediate effects of better information, deliberation contributes to the legitimacy of policy choices and of the underlying political institutions in a number of other ways. It can raise the sense of political autonomy and of the effective fairness of policy choices, enable a better assessment of fellow citizens' motives with respect to a given political choice, and even encourage other-regarding motives on their part (Elster 1995). It may also increase the stability of collective choice by reducing the number of issue dimensions and introducing a greater structure into individual preferences (Johnson and Knight 1994; Dryzek and List

2003). But to have these effects, deliberation must bring about some kind of learning that can produce a change in participants' preferences over choices.[1] At bottom, the transmission, processing, and aggregation of information that forms the basis of individual and collective decision-making is the engine that sets in motion the deliberative wheels.[2]

Although normative theorists have fashioned the research agenda of deliberative democratic theory,[3] that same informational engine is the key motivation behind the now very extensive game-theoretic literature on communication. An emerging body of work in this literature focuses specifically on policymaking in deliberative institutions. One might anticipate, therefore, that there would be a great deal of interaction between the normative and the game-theoretic work on deliberation. Alas, that is not the case. While the game-theoretic studies sometimes explicitly challenge assertions that emerge in the normative literature, they tend to exert little effort in making their results accessible to a considerably less technical normative theory readership and rarely undertake the reconstruction of the normative arguments in a manner consistent with the insights from the game-theoretic models. The influence in the opposite direction is, arguably, even weaker: with very few exceptions, the normative literature on deliberation has, essentially, taken no account of the presence of the game-theoretic work on deliberation and ignores the fundamental incentive problems that surface in nearly all relevant game-theoretic studies.

The dissimilar styles of exposition and the high entry costs are, surely, at least in part responsible for why the two approaches do not see eye-to-eye. But though not irrelevant,

---

[1]This is, arguably, true regardless of whether the implied transformation is of the participants' "primitive" or their "induced" preferences.

[2]See the instructive discussion of causal claims on behalf of deliberation in Fearon (1998).

[3]Throughout, we use the terms "normative theory" and "political philosophy" and their derivatives interchangeably.

these reasons are, not surprisingly, rarely invoked self-consciously. The explicit reasons that bear greater prima facie defensibility are, in our reading, three-fold: (1) it is unclear how to make sense of the analytical/structural relationship between these approaches: what kind of contribution, if any, can each approach make to the pursuit of the agenda set by the other? (2) the communication analyzed by game theorists is of a fundamentally different epistemic type, and the game-theoretic results are, therefore, largely irrelevant to deliberative democratic theory and (3) the game-theoretic approach omits key social and philosophic determinants of deliberation, and its conclusions are therefore also irrelevant because they are an artifact of these omissions. In one form or another these reasons have been called up to justify what we believe is an unfortunate divide between the two approaches. While sometimes useful in maintaining the conversation - itself, ultimately, a rarity - they are also, on close examination, not supportable as reasons for maintaining the status quo course of mutual avoidance.

Our goal in this paper is to sketch what we take to be the most productive and defensible way to think about the relationship between the two approaches and to make, against the critics, the case for the relevance of the existing game-theoretic analysis of deliberation to the development of deliberative democratic theory. Our mode of analysis is, in large part, reconstructive - we suggest ways of thinking about the two approaches that may differ from (though need not be at odds with) the explicit ways in which their respective partisans think about them, and deliberately avoid a sustained survey of the results from the existing work in the interest of focusing on the broad-brush questions of underlying analytical relationships. Since our goal is to reach a broad audience with little or no familiarity with game-theoretic reasoning, we also adopt a non-technical exposition and offer informal self-contained examples to illustrate our arguments.

Decomposing Social Interactions

In this section we lay out the analytical details of our view of the relationship between normative and game-theoretic approaches to deliberation. The discussion in the remainder of the paper addresses a number of key issues related to how and whether the game-theoretic contributions to the analysis of deliberation further the research agenda implied by this view. The next section explores some of the distinctions that have surfaced in the game-theoretic treatments and makes the case that the incentive problems that motivate such treatments persist across the epistemic environments that have been the focus of deliberative democrats. The following section builds on this case and responds to the critics who view the game-theoretic results as an artifact of sins of omission.

Behavior vs. Environment

One might expect that a first step in advancing our understanding of deliberation would entail settling on the precise definition of this concept. Perhaps surprisingly, then, the general label of "deliberation" is used to refer to a broad range of phenomena that are sometimes only tangentially related. A somewhat rough but relatively faithful way of capturing the key underlying differences in formulating definitions of deliberative democracy turns on the difference between *behavior* and *environment*.

For most normative theorists, the focus of the theoretical enterprise is on specifying constraints on individual and group behavior in and around deliberation (see, e.g., Guttman and Thompson 1996; Bohman and Rehg 1997; Dryzek 2000). For this perspective a stylized example is a claim along the lines of "deliberative democracy is a process by which individuals offer justifications, minority interests are not ignored, and political decisions which benefit the group's interest are chosen." Scholarship in this tradition is focused on the ap-

propriate definition of a normative ideal of deliberative democracy - what an ideal practice of deliberative democracy looks like. It advances this goal by presenting a description of what ideal behavior by citizens looks like, including what that behavior looks like in relation to particular social and political institutions and settings. Sometimes such argumentation also concerns outcomes (epistemic, distributive, or other) that may be expected to result from a deliberative democratic practice, but its consideration of outcomes is mainly a by-product of the focus on behavior. The underlying claim is that it is such and such ideal behavior that will bring about such and such outcomes. In short, this scholarship may be understood as treating deliberation and deliberative democracy as behavior - a profile of (normatively defensible) actions and choices on the part of the citizens, or, equivalently, a profile of restrictions on what admissible behavior would be.[4]

In contrast, the approach taken by game theorists involves defining deliberation not as a set of restrictions on behavior but rather a set of restrictions on the environment in which participants interact. In this approach, a stylized example is a claim "deliberative democracy is an institution in which participants have the opportunity to make speeches prior to voting." This claim, then, restricts attention to the decision-making that would be arrived at *in a particular way* (namely, by first exchanging speeches, and then casting ballots based on the judgments that, presumably, reflect the content of prior speech).

The game-theoretic approach involves a three-step process. The first step defines a game, which captures (a) the relevant choices that are understood to be available to the players (in models of deliberation, typically, what messages, if any, could be sent, and what deci-

[4]Of course, the term "behavioral" has many meanings in political science. Some apply this label to scholarship on deliberation that emerges from the social psychology and experimental traditions. See Mendelberg (2002) for a review of this literature that is outside the scope of our arguments in this paper.

sions could be made after the exchange of messages), (b) what the players know about those choices, about each other, and about the deliberative interaction to which they are a party, and finally, (c) how attractive they would perceive the consequences of those choices to be if they knew everything that there was to know about them. The second step specifies a solution concept, which embodies a set of assumptions about the general behavioral agency ascribed to the players in the model. Given the first two steps, the third step is logically entailed: through well-defined techniques of analysis, one can generate predictions about what types of behavior, with respect to the particular choices analyzed in the model, are and are not mutually consistent - that is, are or are not supportable by *equilibria* of the specified game. The key question that motivates the game-theoretic analysis is how policy selection is related to private information and preferences when participants engage in equilibrium behavior. Although this work aims to produce characterizations of behavior, it is important to see that what it has to say about deliberative behavior is, in the sense made clear above, induced by its focus on the nature of the deliberative environment. The analysis it delivers is the analysis of the properties of that environment, and it contributes to the study of deliberation insofar as that environment captures the essential institutional features of deliberative democracy. In short, this work treats deliberative democracy as an environment.

The distinction between treating deliberation as behavior and treating it as an environment is a conceptual divide, rather than a sociological one, separating normative from formal theorists. (In this sense, our usage of the term "game-theoretic," as opposed to "formal" is revealing. The critical contrast is not between formalized theory and non-formalized theory, but between the game-theoretic focus on the deliberative environment and the normative behavioral focus on the behavior without inducing it from the environment.) As a methodological matter, a sharp distinction between these foci pertains also to the way in which the two approaches substantiate positive statements. Work on deliberation as environment

typically posits *an external set of behavioral assumptions* that have implications for how collective bodies behave in any environment. This enables the deductive determination of what deliberation will look like in a particular environment subject to the condition that the external assumptions are true. This approach suggests two directions for analysis – varying the external assumptions about how individuals behave in a particular environment (Hafer and Landa 2003) and varying the description of the environment, holding fixed the external assumptions (Austen-Smith and Feddersen 2005a,b; Meirowitz 2003, 2005; Hafer and Landa 2006). As the difference between these directions indicates, the analysis of deliberation as environment does not require that the external assumptions conform to a particular *a priori* fixed notion of rationality.

In contrast, scholarship in the behavior approach is not subject to explicit rules about "how people operate"; it is thus harder to evaluate any given account that approach may generate. Whereas the game-theoretic/environmental approach has an agreed upon "machine" (or, more accurately, a small set of "machines") for relating descriptions of the environment to descriptions of behavior, the deliberation as behavior approach lacks such a device, thus making it difficult for it to evaluate behavior-based claims about such relations. Each behavior-focused account of what deliberation means includes a potentially novel description of the rules of behavior and the rules for drawing inferences. The point is not that this approach necessarily does a worse job capturing reality than the game-theoretic one. It is, rather, that it has not equipped itself with an epistemic mechanism for discriminating between different accounts that instantiate it.

While this contrast underscores the relative appeal of the game-theoretic approach for generating comparisons about how different descriptions of the environment might influence the type of discourse and policymaking, this approach is not a panacea. The game theorists should, no doubt, debate which behavioral agency is appropriate and which environments

are relevant. As we argue next, moreover, the range of things they can say about deliberation is also intimately linked to the normative insights about equilibrium behavior.

Democratic Ideals and Game-Theoretic Equilibria

The aim of deliberative democratic theory is to develop a set of claims about the properties of deliberative practice in democracies, including claims that buttress a normative justification of institutional and policy choices that give prominence to deliberative social interactions. The contrast between the focus on deliberation as behavior and on deliberation as an environment highlights what we believe is a fundamental difference between these approaches. The insularity associated with this difference is not, however, sustainable if the corresponding traditions aim to advance scholarship on deliberative democracy. To see why, it is helpful to consider what we hope to learn from the analysis of social interactions involving deliberation, and to ask how these respective research traditions contribute to this overarching project.

As we noted above, the primary source of the deliberation as behavior approach has been the work of normative theorists. One may characterize a key contribution of this work as the articulation of axioms that are definitive of the democratic ideal and of its various conceptions - e.g., participatory, deliberative, epistemic, representative, direct, etc. Some of these axioms are procedural - i.e., descriptive of the process of deliberation and decision-making, such as honest and open-minded participation, etc. (e.g., Cohen 1997; Gutmann and Thompson 1996). Other axioms are best understood as consequentialist - e.g., collective decision-making selects a policy that is (more) correct, or carries with it minimal expected deviation from the policy that would have been chosen by a single omniscient decision-maker with complete information (Estlund 1997).

Of course, normative theorists do not merely define the axioms, they aim also to make

the case that, in relation to those axioms, some conceptions of democracy and some social institutions are better than others at capturing the democratic ideal. It is in connection with this step that we argue for the importance of a give-and-take with the game-theoretic analysis of deliberative environments.

In order for communication to do more than just allow participants to coordinate on a particular choice (that is, if deliberation is about convincing and/or being convinced by one's interlocutors), one or more participants must be uncertain about some aspect of policy choice.[5] The game-theoretic approach to modeling situations with such uncertainty is to assume that some, and possibly all, participants are endowed with pieces of information that their counterparts do not know but would find relevant to the decision at hand.[6] Other participants know that the first participant might have some privileged information but do not know what that information is. An instructive analogy may be drawn to a simple version of the game of poker - a strategic environment that is, undoubtedly, considerably more trivial than that of many actual deliberative interactions but that, nonetheless, shares some of their salient features. Imagine a group of card players, and let some of each player's cards be dealt face up and some face down. The face up cards are publicly observed, while the face down cards are privately observed. Each player knows that every other player has seen her own face-down cards even though she does not know them herself. Each player can also form beliefs about the other face down cards. The beliefs that a particular participant generates are based on the cards available in the standard deck, the observed cards that were dealt

[5]See, though, Calvert (2006) for a model of coordination via cheap-talk deliberation that does not have policy uncertainty.

[6]This uncertainty can range from the effects of policy choices, and the identity of a true state of the world, to the logical (most efficient) way of organizing other information that is already available to the participants.

faced up and her own cards that were dealt face down. The first two pieces of information are available to all players, but the last is known only to the holder of those cards.

Similarly, in deliberative environments with uncertainty, although some participants are uncertain about elements of information or arguments that may be relevant to them (e.g. about the consequences of policies, the preferences or tastes of participants, valid or believable arguments that would speak in favor of certain decisions), they may be expected to develop beliefs about these elements based on what they know or find convincing. Those preliminary beliefs enable them to make educated decisions in the absence of further information and to make sense of the signals or cues they may receive from others, but on their own, those beliefs are "noisy." A key issue in game-theoretic models of policy-making is whether it is reasonable to expect those participants who possess valuable private information to reveal it to others, and whether those others have good reasons to believe it. In the poker example such an expectation is unreasonable. We would not expect a player to actually reveal that she has nothing, or that she has a royal flush. In fact it is this aspect of poker that makes it entertaining to watch and play. We observe small-talk and gestures, but the savvy player will not take such communication at its face value - not the least, because she knows that competitive players go to great lengths to make sure that non-verbal communication does not betray them.

The equilibrium notions that game theorists use to analyze such interactions are nothing more than descriptions of how the participants play the game. Such a description must satisfy the condition that, when the players share a common conjecture about how the game is being played, no player (conditional on any particular realization of her private information) has an incentive to deviate from the behavior that that description ascribes to her. In practice an equilibrium concept may be a bit more complicated because it also specifies the inferences that participants should make when particular speeches are observed.

These inferences depend on the conjectured behavior (e.g., a conjecture that players with straights tend to be very quiet), the underlying environment (e.g., the distribution of Aces in a standard deck of cards), and the assumption of behavioral agency being maintained. In other words while participants may possess uncertainty about what others know, an equilibrium consists of a conjecture about the rules governing the behavior of individuals (e.g., always tell the truth, or tell all and only the truth under conditions x, y, or z and lie in some specified way otherwise), and in an equilibrium, participants draw inferences from the speeches that are made, using that conjecture. Given the equilibrium conjecture, when participant A hears participant B say something like "going to war is a terrible idea," participant A can form beliefs about what participant B knows about the relevant policy alternatives. An equilibrium corresponding to a deliberative environment must satisfy three conditions: (1) Given the way that participants are forming beliefs based on communication, participants vote for the policies that they think they like best. (2) Given the ways that participants are behaving in the policy selection stage (which is dependent on how they are forming these beliefs based on communication) all of the participants have an incentive to communicate in the manner conjectured. (3) Given the way that participants are conjectured to behave at the communication stage, the beliefs that are formed are consistent with the postulated process.[7] Taken together, these criteria of equilibrium behavior amount to the requirement of what may be called *strategic behavioral consistency* - the mutual consistency of individual behavior in a game-theoretic equilibrium.

To make the implications of this requirement a bit clearer, we return to the poker example,

[7]Examples include the standard use of Bayes Rule, in which case the equilibrium is sometimes termed the Perfect Bayesian Equilibrium. Alternatively, belief revision may allow for the possibility of inefficient updating, as in the models in Hafer and Landa (2003, 2005).

and ask what types of behavior can occur in equilibrium. Is it possible for there to be an equilibrium in which participants truthfully announce their face down cards to each other? It is not surprising that the answer is no. The benefits of successful bluffing insure that such communication is not credible. Less obvious is the claim that it is not possible in equilibrium for all players with good hands to be quiet and all players with bad hands to be chatty. In such an equilibrium a player with a bad hand might have an overwhelming incentive to refrain from chattiness, and fool her opponents into believing that she had a good hand. Also players with good hands may try to chat in order to up the betting. As we will see below, this type of reasoning has been applied to the study of deliberative policy-making.

The notion of game-theoretic equilibrium is a predictive device that describes what to expect from a social interaction in a given environment. For our purposes, a particularly useful way of thinking about equilibria is in more hypothetical terms of *implementable collective choice functions* - relationships between policy selection and the participants' private information and preferences that are consistent with mutually rational or equilibrium play. In focusing on equilibria, game theorists, in effect, privilege implementable collective choice functions associated with corresponding descriptions of the environment over those that cannot be implemented without altering some of the underlying elements of that environment or the general behavioral agency posited in the model.[8] If for a particular environment and model of behavioral agency, there are no equilibria in which participants share all of their information, then a game-theorist is suspicious of the expectations that participants will be

[8]Some game-theoretic work (e.g., Austen-Smith and Feddersen 2005a, Meirowitz 2003, Levy 2006, Stasavage 2006) treats a very stylized game generating a precise description of "equilibrium play." Others (e.g., Gerardi and Yariv 2004, Austen-Smith and Feddersen 2005b, Meirowitz 2005) treat large classes of particular games, but generate somewhat less precise characterizations of deliberative behavior within a given entailed game.

truthful in deliberative settings that match that environment.

The axioms that form the basis of normative theorists' descriptions of democratic ideals may be seen as the codifications of expectations, both with respect to the aspects of deliberative environments and to the behavioral choice in deliberation and voting. In seeking to satisfy these axioms together, we hope for what may be called *axiomatic consistency*. A key claim of the game-theoretic approach may, then, be put as follows: an appropriately specified game-theoretic model allows us to gauge axiomatic consistency by ascertaining the satisfaction of strategic behavioral consistency. In effect, by analyzing the properties of the equilibria of the relevant game-theoretic model, we obtain propositions about the compatibility of various behavioral expectations with each other and with the particular aspects of the deliberative environment that define the game. In the equilibrium analysis, this test of axiomatic consistency may happen ex post: an axiom that is thought to ensure a particular consequence may, through the equilibrium analysis of deliberative environments that vary with respect to it, sometimes be discovered to imply altogether different, unanticipated, consequences as well (or, put differently, be compatible or incompatible with a particular set of ex-post expectations or axioms). In this way, recent game-theoretic work on deliberation has called into question the expectations associated with such axioms as the preference for participation and diversity of deliberative bodies (Meirowitz 2003), the equality of opportunity to make one's arguments heard (Hafer and Landa 2003), and the requirement of consensus in collective decision-making (Gerardi and Yariv 2004; Austen-Smith and Feddersen 2005b).[9]

The game-theoretic analysis of deliberation as environment may, thus, be seen as providing a critical independent test to aid in adjudicating competing claims about deliberative

[9] While the game-theoretic work proceeds in a different idiom than the normative literature, it is clear that the game-theorists are not always unwilling to confront its conclusions.

behavior. When two theories yield distinct descriptions of deliberation (as behavior), they naturally raise the question of how one is to determine which description is privileged. Because the theories that yield these different descriptions tend also to rely for their support on differing philosophical criteria, if not differing general philosophical outlooks, the adjudication between them by reference to those criteria is rarely resolute. In contrast, pragmatic criteria like empirical support and implementability within the underlying environment provide adjudicatory frameworks that cut across the philosophical divides.

The Three-Step Deliberative Democratic Theory

To the extent that democratic theory has practical aspirations, it seems difficult to argue that the mutual inconsistency of axioms that characterize normative conceptions of democracy is irrelevant. This suggests the importance of developing scholarship on deliberative democracy that is responsive to the considerations of consistency or implementability offered by the game-theoretic analysis. In our view, scholarship of this form may be instructively conceived as consisting of the following three steps. The first step corresponds to the formulation of axioms, both procedural and consequentialist, in relation to the underlying political environment (the environment that, arguably, captures the structural features of a given conception of democracy - e.g., for deliberative democracy, the environment described by majority decision-making preceded by unforced communication between freely associating voters). The second step entails the analysis of axiomatic consistency within a corresponding game-theoretic model, including the consistency of the proposed axioms with the behavioral agency that is thought to characterize the agents operating in that deliberative environment. The third step in the construction of deliberative democratic theory closes the loop: it calls for a review of the normative conceptions and axioms with which the process began and sanctioning trade-offs where they are necessary. Whereas the first step may be largely

the domain of normative theorists, and the second step that of game theorists, the third step presupposes a conversation between these research traditions:[10] if not all of the goods things (expressed by the relevant axioms) can go together, and not all of those good things have the consequences that we thought them to have, what trade-offs between axioms are justifiable?[11]

One aspect of what game-theoretic analysis brings to this conversation deserves special emphasis. The assessment of the mutual compatibility of behavioral axioms allows us to conduct comparisons of different deliberative environments by comparing properties of their equilibria. Because environments are, in part, defined by particular institutions instantiating them, we can, proceeding in this fashion, arrive at a ranking of institutions in relation to various normative properties. In so doing, we can provide support for the normative arguments regarding institutional choice.

A central intuition behind this type of work - its current examples include considerations

---

[10]Johnson (1993) urges a somewhat different conversation between game theory and deliberative (critical) theorists. As he notes, the problems of indeterminacy that arise in game-theoretic contexts - mainly as a consequence of strategic uncertainty in the multiple-equilibrium environments - suggest that strategic action is insufficient as the source of social coordination and that the missing elements may be supplied by binding speech acts. (See Heath 2001 for further development of this point.) Recent work on non cheap-talk communication that we discuss below is, in part, motivated by the same concern. However, the issues raised by Johnson's argument concern more directly the general question of the shape of the social theory of action, which is far broader than the aims and format of this paper allows us to consider.

[11]With a few exceptions (e.g., Johnson and Knight 2005; Landa 2005), work on the third step appears to be nonexistent.

of voting rules, types of participation, speaking rights, and non-policy side-payments or incentives - is the idea that institutions have consequences for individual choices and, through these choices, for social outcomes. Because the outcomes are likely to differ with respect to their epistemic and welfare properties, a normative theory of deliberative democracy must be a theory of institutions. A failure to recognize this point is responsible for the view that the normative arguments of deliberative democrats are somehow inconsistent with the notion of individual rationality. According to this view, most recently advanced by Posner (2004), if individuals are choosing the nature of their deliberative engagement rationally, then their choices imply that any further argument in favor of deliberation must, in effect, council behavior that is contrary to rationality.

It is critical to see that this view rests on a fundamental mistake about the determinants of social outcomes. Different institutions provide individuals with different incentives by, inter alia, allocating resources that affect the size and the nature of one's potential audience (such as electoral campaign subsidies, time in front of the microphone in a committee meeting, etc.), changing which voter is pivotal (by changing the voting rules, the degree of centralization in collective decision-making, etc.), distributing decision-making authority across levels of government, etc. Although the incentives these institutional choices create are not always immediate to outside observers, individuals that are subject to them can be expected to recognize them over time and adapt their choices regarding deliberation accordingly. To evaluate institutions with respect to properties of deliberative outcomes is to ask whether, relative to other institutions, they create incentives such that, in the aggregate, the outcomes they give rise to are at or near the implementable optimum. (For extended philosophical arguments to this effect, see, for example, Parfit 1984 and Hardin 1988).

This argument underlies both a causal claim about institutions and a methodological recommendation in connection with the normative arguments regarding deliberation. The

former is by now clear: institutions affect behavioral choices and, through them, the properties of deliberative outcomes. The latter is, arguably, implicit in the causal claim: normative arguments regarding institutional choice must treat deliberative behavior as, inter alia, dependent on the institution. If they do not, there is no reason to believe that the deliberative and policy outcomes will, in fact, have properties that are consistent with the expectations associated with the endorsed institutions.

Incentives in Deliberative Environments

Discourse, Policy Making, and Incentives

In comparing deliberative environments, an immediate question to confront is whether the focus of analysis is on discussion and debate or on discussion and debate by participants that are directly charged with making policy decisions. As we discuss in greater detail below, recent findings by game theorists suggest that this distinction has critical consequences for what may be expected from the deliberative process. When communication precedes policy-making, participants will have incentives to misrepresent or withhold information unless their underlying values or preferences are commonly known to be quite similar. In contrast, if one considers a deliberative environment in which discussion is not followed by policy-making, and participants care only about whether they themselves arrive at the most defensible judgments, it can be easier to sustain informative truthful debate. This difference in incentives is important not simply because game-theorists care about strategic behavior, but because, as we emphasized in the preceding, the institutions of policy-making may be expected to have consequences for individual choices, including individual choice in the debate prior to the decision-making. Whether the debate is followed by policy-making becomes a critical factor in ascertaining what behavior may be expected in the debate.

This conclusion underscores the importance of clarity with respect to how the analyzed discourse fits into the larger spectrum of political interactions. Because the development of a formal model forces an explicit description of the environment, it is clear that positive theorists have been almost exclusively (Glazer and Rubinstein 2001 is one of very few exceptions) interested in debate that precedes policy-making and have not focused on discourses without some subsequent consequentialist action. By contrast, the distinction between discourse and deliberative democratic policy-making is not always clear in the work of normative and social psychology scholars. Indeed, while some scholars theorize deliberation as an institutionalized part of an explicit electoral process, they seek supporting evidence in the analysis of deliberative practices in the contexts in which the relationship to the policy-making decisions is either absent or ambiguous (e.g., Fishkin 1991, Barabas 2004).

We believe that deliberation in politics is, in fact, best understood as leading to policy choices. Whether those choices are collective (e.g., whether to protect the right to obtain abortions) or individual (whether to obtain an abortion), they have externalities - that is, they affect how others perceive their welfare and create incentives for them to influence the selection of more favorable alternatives. To the extent that deliberation is consequential for what policies are chosen, these incentives may be expected to affect individual behavior. For example, the tone of a private debate about abortion in a setting divorced from any actual abortions is likely to be quite different from the tone of debate between an opponent of abortions and a person contemplating an abortion. In the remainder of this section we focus on the distinct environments in which participants debate and ultimately select a policy, and consider some of the diversity of incentives that can surface in them.

Common Values, Private Values, and "Cheap Talk"

Consider another parlor game which might be closer to policy-making than the poker game described above. A group of individuals must decide as a group whether to bet on a particular die coming up on an even or odd number. Each group member gets a share of the groups winnings. The individuals in the group are uncertain about the odds that the die comes up odd or even (i.e., how fair or unfair it is). Prior to betting each member gets to privately observe one toss of the die. The group members then assemble and are given the opportunity to talk and vote (say the voting is under majority rule) over which way the group should bet. Assume that the group members all want to maximize their earnings and that all members of the group know this. In this experiment we would expect the group discussion to be much more informative than in a poker game. In fact, we might expect them to be as informative as possible, truthfully revealing what they observed and discussing which bet is best. It is possible to show that these expectations are consistent with the formal requirements of equilibrium play.[12]

In reaching this conclusion about equilibrium play, we made a critical assumption: it is commonly known that all members of the group get a share of the group's earnings and so are interested in maximizing it. Game theorists are particularly attuned to assumptions like this. Why is such specificity needed? Suppose that one of the participants, say player 1, receives a positive payoff if the group's bet is incorrect and a negative payoff if the groups bet is correct (that is, 1's incentives are like those of the "house"). In this case player 1 might have an incentive to mis-report the outcome of the toss that she alone observed in the hopes of leading the group to the wrong decision. In a setting in which the other players do

[12]It should be noted, that other types of behavior, including speeches that are entirely uninformative are also consistent with reasonable equilibrium concepts.

not know that player 1 has this conflict of interest 1's report might be believed. If, however, all of the players know that player 1 has a different motivation, then there cannot be an equilibrium in which she is taken at her word. If player 1 could say something that would lead the other participants to believe that a particular outcome were more likely then she would want to send this type of message whenever she thought that outcome was unlikely. In game-theoretic models it matters not just what preferences one has but also the beliefs that other players have about one's preferences.

A group in which it is known that all participants have the same, or common preferences and a group in which it is known that participants have opposing preferences represent polar cases of a spectrum. Game theorists typically use the term *common values* to describe the former end, and *private values* to describe the latter end. (The term interdependent values is used to describe the remaining cases.) Corresponding to these distinctions between groups, it is convenient to distinguish between the incentives faced by their respective members as common values problems and non-common values problems. The good news is that in common values problems honesty is a reasonable expectation. More precisely, there are equilibria in which participants truthfully reveal what they know and the voting reflects all of the available information. The bad news, however, is that in non-common values problems, expectations of honesty may be unwarranted (Meirowitz 2003 and 2005).

Why should a deliberative democrat be concerned with the distinction between private and common values? One of the central issues of contention among deliberative democrats is the expectation of consensus following deliberation. Jürgen Habermas (e.g., 1996) has argued for the reasonableness of such an expectation - in particular, that, if participants could argue indefinitely, they would converge on the same judgments. The preceding discussion suggests that given rational agency, that expectation is either wrong or its validity hinges on the assumption of common values. It is important to emphasize that common values does not

mean merely identical primitive preferences. It is far more demanding: it requires that the interlocutors have common knowledge of this identity (that is, that each of them knows that the others know that he knows ... that primitive preferences are identical). Thus, the very fact that the scholars of deliberative democracy disagree on this point (see, e.g., Bohman and Rehg 1996) strongly suggests that the assumption of common values is untenable (even if participants *actually* have the same preferences). To be slightly tongue-and-cheek, we can say that any collection of individuals that includes the published normative theorists does not have common knowledge of common values.

If the paradigmatic deliberative interaction involves non-common values, then the above discussion suggests the importance of taking seriously the incentives faced by the participants with respect to their deliberative choices: what, if any, information and arguments to share with others, whether to engage in misrepresentation, and how to interpret the communications of others. In particular, this suggests the importance of two questions about deliberative environments: when is it the case that there are equilibria in which participants are truthful in their speech-making? and what are the properties of the equilibrium policy selections? Once the analyst can answer these types of questions, she can then go on to answer questions like, given an agreed notion of desirability, what is the best type of environment?

Incentives in Richer Deliberative Environments

Both in the poker and in the odd-even die examples, speaking imposes no costs on the speaker, and that means that the listener or observer cannot determine which speeches are truthful and which are not by cuing off the speaker's cost of making them. The only thing that they have to go on is the word of the speaker and their own understanding of the nature of the interaction. Deliberation in which participants' statements may be construed in this fashion - as *cheap talk* - is a frequent feature of social and political interactions in democratic

institutions.[13] The problems of (in-)sincere speech, and in consequence, of under-informed post-deliberative decision-making, that arise in the environments without common values may be commonplace in these settings. Still, cheap-talk arguments are not the only kinds of arguments that could be made in deliberation, and this naturally raises the question of the scope of deliberative interactions with the strategic problems that we described above.

In cheap-talk deliberation, the mechanism for inference relies on the listeners's ability to pin down the determinants of utterances. That is, the listener must ask the following question: given what is known about the nature of the deliberative interaction, including the preferences and the beliefs of the speakers, what kind of private information could the speaker have that would be consistent with the observed speech, given that the speaker is making rational decisions about what to say (that is, understanding how her speech will likely be perceived)? In order for an equilibrium to involve learning, the listener(s) must correctly believe that the speaker(s) would like the listener(s) to learn what is learned. There is, however, another mechanism for inference, as well, and it corresponds to another type of argument. The listener may determine the validity of speaker's statements through the consideration of undeniable evidence: for example, somehow being allowed to catch a glimpse of the opponent's cards, in the poker game.

Arguments in deliberation may display different positive degrees of direct (or intrinsic) provability. An argument could be *partially provable* in the sense that only a part of it is verifiable, leaving some residual uncertainty about its validity. The argument could also be

[13]This usage, which has become standard, combines two logically distinct attributes: costless and non-verifiable (non-provable) communication. In this section we consider costless but provable communication. This type of communication can be considered a form of cheap-talk in terms of the first attribute alone. Our reference to it as an alternative to cheap talk is, however, in keeping with the standard usage.

fully provable - that is, we could know the full merits of the argument (though it may also happen that we could obtain this proof of validity with some positive probability that is short of certainty). The provability of arguments may stem from their logical persuasiveness or unfalsifiable hard evidence. The possibility of argument provability may at first suggest that concerns about the strategic dissembling that are rampant in the cheap-talk environment with private values may be irrelevant. This conjecture appears in Cohen (1998) and Mackie (1998). Alas, in connection with partially and fully provable arguments, private values raise another set of strategic concerns. Verifying the validity of arguments is typically costly - it requires an investment into becoming informed (either literally paying for corroboration, or incurring the direct cost of acquiring expertise on one's own, or incurring the opportunity cost of listening and processing when a claim is being justified with a potentially provable argument). The existence of such costs means that a speaker with a primitive preference that is potentially different from that of her audience will sometimes want to make arguments that are provably wrong in the hope of passing them off as (possibly) right ones. In such cases, the speaker may rationally anticipate that the listener will not incur the cost of verifying the argument, taking it as true with some positive probability. Alternatively, if verifying the argument would require sacrificing resources that may otherwise be spent in ways contrary to the speaker's preference, the speaker may have a further interest in making provably wrong arguments in the hope that the listener incurs such a cost (Landa 2005).[14]

More striking, however, is the fact that even when arguments are instantaneously and

[14]It should be noted that in the case of non common values and costly monitoring if in fact equilibria with full revelation exist then monitoring must occur with positive probability. This monitoring itself represents a form of inefficiency –a point missed by the proponents of the claim that monitoring solves the incentive problems present in cheap-talk models (most notably Mackie, 1998).

costlessly verifiable, incentive problems surface. The following example provides an illustration. Suppose a group of committee members are deciding whether to support Jill, a candidate for city office. A majority of the committee members are concerned about Jill's integrity but like her policy ideas. They will support her unless they have clear evidence of improprieties. Jack, a well connected member of the committee happens to know Jill well, and thus might have some additional information about Jill. It is possible that he is aware of indiscretions and can provide documentation to the committees. It is also possible that Jack possesses no additional information about Jill's past. Suppose that Jack is exceptional in his tendency to not let a person's past influence his assessment of their ability to perform in the future; Jack believes that regardless of her past Jill is the best person for the job. What should Jack do if he actually does have evidence of past indiscretions? What will Jack do? If Jack places enough weight on his belief that Jill should win he might choose not to inform the other committee members about Jill's past indiscretions. On the other hand, if Jack puts enough weight on the process or on being forthcoming, then he will provide the evidence. In this non-cheap talk setting, Jack's communication choices will depend on the alignment between his preferences with those of the other committee members; if he places enough weight on his belief that Jill is best for the job he will strategically refrain from providing information that is valuable to the other committee members. Knowing that Jack might have an incentive problem, how will the other members interpret silence by Jack? It cannot be taken as evidence that he does not know of any indiscretions by Jill. So if, in fact, Jack knows that Jill has not been guilty of any indiscretions, the incentive problem prevents him from conveying this information to the other members. The other committee members cannot know if Jack's silence stems from a preference alignment problem or the fact that he knows of no indiscretions, and thus the committee looses the ability to use Jack's expertise about Jill to its fullest extent. Driving the incentive problem in this discussion is the fact

that, even when arguments are provable, the speaker has a choice about whether to make the argument. There is no slight of hand here; a speaker can refrain from providing a provable argument about a question as long as there is uncertainty about whether she knows the answer.

Parallel to the distinction between common and private values, there may exist one between common and private *veridicality.* Common veridicality characterizes the environment in which it is commonly known that all participants would agree as to which of the arguments are persuasive and which are not, whereas private veridicality corresponds to the case in which it is commonly known that such an agreement may not exist. Although we tend to associate positive provability with common veridicality (e.g., if Fermat's last theorem is true, it is true for everyone), deliberations in politics often combine provability with private veridicality (e.g., in a room full of people with varying levels of mathematical training a proof of Fermat's last theorem may be compelling to some but not others). The intuition for this possibility is that if citizens' moral commitments or values are sufficiently different, this difference may be expected to affect what arguments they find persuasive (e.g., for and against legalization of gay marriage, for and against racial profiling, etc.). In such cases, arguments that combine provability with private veridicality often proceed by articulating logical consequences of claims that some part, but typically not the whole, of the audience accepts as true but, for whatever reason, mistakenly considered irrelevant to the issue at hand.

In principle, whether the environment is one of common or private value may be interpreted to be a function of what arguments participants in the deliberation may find persuasive - that is, of the underlying veridicality. Often, however, it is more useful to think of common/private veridicality and common/private values as distinct. Thus, we often speak of "good-faith disagreements," meaning something like the cases of common values paired

with private veridicality. In contrast, debate between experts with vested interests may be thought to involve the opposite pairing of private values and common veridicality.

In the environment with common values and common veridicality, deliberation may be seen as a purely cooperative endeavor: setting aside the costs of deliberative engagement, individuals may be expected to seek to convince their interlocutors by sharing with them everything they know. Private veridicality changes this incentive. If deliberation is, indeed, followed by policy choice, then even in the cases of would-be good-faith disagreements, deliberators have incentives to avoid making a (dis-)provable argument if doing so may convince the listener that she is the sort of agent who finds that sort of argument unpersuasive - and make a policy choice that reflects that. Even when agents have aligned preferences, private veridicality may result in incentives to withhold information.

The following example provides an illustration. Suppose that Jill is uncertain in her position on abortion rights: she leans against them, but is not yet convinced that she has heard the most persuasive arguments that would support her position. She also recognizes that she may not have heard the most persuasive arguments against her position. Suppose, further, that Jill finds herself in a discussion with Jack, who she has good reason to believe is one of the most thoughtful critics of abortion, and Jack makes to her a series of arguments, none of which she finds persuasive. How should this interaction affect Jill's beliefs about the justifiability of abortion rights? She should conclude that abortion rights are more defensible than she had previously thought; here is one of the most compelling critics and his arguments are not particularly persuasive. Knowing that Jill would otherwise be likely to support anti-abortion politicians, Jack may be reluctant to make his arguments to her. In short, in the case of private veridicality, the speaker will often have an incentive to avoid making provable arguments (Hafer and Landa 2006), and in equilibrium, the astute listeners will discount the arguments she hears, thinking them uninformative (that is, speaking to

neither the truth nor the falsity of the relevant claims).

The preceding discussion suggests that the kinds of strategic incentive problems that arise in the cheap-talk environment arise in other informational environments as well. Because criteria of deliberative performance in democratic societies must surely include the extent to which the post-deliberative judgments of citizens are educated by the deliberative process - that is, the extent to which they are close to what they would be if such a process resulted in the articulation of all the relevant information and arguments - the presence of incentives to refrain from making provable and fully informative non-provable arguments underscores the importance of equilibrium analysis for normative theories of institutional choice.

Are Game-theoretic Conclusions an Artifact?

While the recent work featuring formal models of deliberation represents a "pushing of the envelope," the possibility that non-common values might lead to incentives to mislead others has been a prominent fixture in the economics literature since the seminal work by Crawford and Sobel (1982), with applications to the politics of debate appearing in the mainstream political science literature beginning with Austen-Smith (1990). Despite the appearance of a number of formal studies exploring related ideas in the intervening time, normative theorists have largely dismissed these concerns. In the overwhelming majority of cases, this dismissal is "unspoken" - simply ignoring the existence of the parallel research tradition. A more constructive engagement, urged by Johnson (1993) is, to some extent, taken up by Heath (2001), Cohen (1998), and Mackie (1998).

The arguments offered to justify the dismissal of game-theoretic conclusions about incentives in deliberation rest on the observation that the models from which the incentive problems derive are too narrow to capture the entire range of aspects of deliberative envi-

ronments that are relevant to actual participants; the presumption is that once these missing aspects are accounted for, the game-theoretic conclusions about deliberation will lose their bite. Six such arguments - three concerning particular aspects of the deliberative environment, and another three concerning the nature of deliberative behavior - can be usefully distinguished. We consider these arguments and conclude that they are either insufficient to warrant the skepticism of the game-theoretic conclusions or rely on speculations that have failed to withstand close scrutiny.[15]

Strategic Deliberative Environments

The first argument, articulated by Mackie (1998, pp. 84-5), proceeds by observing that participants will care not just about the policy-making of today but also about that of tomorrow. As such, while deception today might result in a more desirable policy outcome it will lead to a reputational hit or other punishment in the future. The problems of information revelation raised in the game-theoretic models are thus, to a considerable extent, an artifact of failing to take into account the long-term nature of the interaction.

While prima facie plausible, this conclusion runs into problems on closer scrutiny. The first such problem is that, while it is true that repeated play may enlarge the set of types of behavior that are consistent with equilibria (a paraphrase of the popular folk-theorems), it is typically the case that equilibria with dishonesty will survive. Thus repetition may introduce the possibility of better information aggregation, but it cannot rule out the possibility of poor aggregation. (As an analog, note that repeated play of the Prisoners' Dilemma does not eradicate the "bad" equilibria, it just introduces the possibility that there are also "good"

[15] We set aside objections to the theory of rational choice as such - a discussion that is outside the scope of the present paper.

equilibria.) A far more striking problem with this argument is that although repetition makes reputational concerns possible, in order to sanction a participant that behaves poorly, the other participants must be able to figure out that the participant behaved poorly. In other words, sanctioning is possible only if it is possible to determine when to sanction. Our first Jack and Jill example above illustrates the point. Since only Jack knows whether he has an argument to make, it is not reasonable for others to punish him if he does not make it. And such a punishment is precisely what is needed to overcome Jack's incentive problem. More generally, external incentives can only solve the incentive problem when (1) such incentives have sufficient magnitude and (2) they can feasibly depend on whether a participant reveals the private information that she should have. Our two examples suggest that the latter requirement can be challenging. While the speculation on the positive effects of repetition on sincere speech has some bite, turning the speculation into an argument requires precisely the kind of detailed analysis of incentives that formal theorists have been engaging in.[16] Proceeding with just such an analysis, Morris (2001) considers a policy-maker seeking advice from an advisor on a policy decision and shows that the shadow of the future can actually create incentives for dishonesty. The following motivating example captures the intuition:

> Consider the plight of an informed social scientist advising an uninformed policy maker on the merits of affirmative action by race. If the social scientist were racist, she would oppose affirmative action. In fact she is not racist, but she has come to the conclusion that affirmative action is an ill-conceived policy to address racism. The policy maker is not racist, but since he believes that

[16] Meirowitz (2005) focuses on the relevance of external incentives and isolates conditions under which the incentive problems can be resolved by external incentives.

> there is a high probability that the social scientist is not racist, he would take an anti-affirmative action recommendation seriously and adjust government policy accordingly. But an anti-affirmative action recommendation would increase the probability that the policy maker believes the social scientist to be racist. If the social scientist is sufficiently concerned about being perceived to be racist, she will have an incentive to lie and recommend affirmative action. But this being the case, she would not be believed even if she sincerely believed in affirmative action and recommended it. Either way, the social scientist's socially valuable information is lost (p. 231-232).

The logic behind this example is particularly damaging to the reputational argument because it is precisely the advisor's concern for her reputation in the eyes of the policy-maker that furnishes her with an incentive to lie, even though both players have, in fact, the same preferences.

The second argument against the relevance of the game-theoretic studies of deliberation is that these studies restrict their attention to a single speaker. With multiple speakers, Mackie (1998) argues, the incentive problems will go away. This speculation about the consequences of multiple senders and/or receivers has been shown to fail systematically across the range of the environments analyzed in the game-theoretic work (starting with Crawford and Sobel's (1982) classic cheap-talk setting, which allows for many senders). To be sure, this speculation is not without some merit. Meirowitz (2005, 2006) shows that under some strong conditions, having multiple participants can help. In particular, if everything that is known by one participant is known by at least two other participants and this is common knowledge, then there are equilibria that fully aggregate the information–as well as equilibria that do not aggregate the information. However, these "nice" results are restricted to special circumstances, and Meirowitz (2003) and Hafer and Landa (2005) demonstrate

that having more participants may, in some environments, not only fail to help, but, in fact, be detrimental for eliciting informative speech. In short, the multiplicity of participants is not a blanket panacea.

The third argument holds that a fundamental shortcoming of the formal analyses of deliberation is their reliance on the "cheap-talk" technology (Cohen 1998; Mackie 1998). We agree that, while much of the formal work analyzes deliberation as cheap-talk, many interesting political applications involve a richer technology. But, as the Jack and Jill examples of the previous section demonstrate, cheap-talk signaling is not a necessary condition for the existence of incentive problems in information-revelation and argument-making. When their arguments are provable, participants will have incentives not to share their information if they anticipate that the deliberative environment is one of private veridicality. The incentive to lie in the "cheap-talk" setting has a counterpart as an incentive to refrain from providing arguments in settings with provability (Hafer and Landa 2003, 2006). What matters is the existence of some form of underlying disagreement - whether directly interest-based or epistemic. When such disagreement is present, the incentives problems may be expected to persist.

For each of these three criticisms, it is possible to construct a model that incorporates the relevant extension of the standard sender-receiver environment in which deliberation can turn out well. But in none of these extensions is it ever the case that deliberation will necessarily turn out well. Rather, one has to impose considerable restrictions on how the particular concern is incorporated in order to generate "nice" results, and it is typically the case that other equilibria with less desirable properties will exist as well.

More broadly, both this conclusion and the critiques of the game-theoretic deliberative environments that occasion it highlight two key methodological points. First, appeals to robustness issues as a way of dismissing inconvenient concerns about incentives represent an

implicit concession that behavior depends on the particulars of the institution. Because a key value of the formal enterprise is that it facilitates robustness checking, the appeal for robustness checks by normative theorists is an implicit argument in favor of continued game-theoretic work on deliberation. In effect, by observing that the particulars of the model can drive the inferences about behavior, critics of the game-theoretic approach encourage the institutional analysis of deliberative democracy that we are urging in this paper. The key point of difference is that they implicitly assert that the "right" robust models will not suffer from incentive problems - an assertion that is broadly refuted by the body of game-theoretic work that incorporates the mentioned extensions. This outcome underscores the second methodological point: it is insufficient to point to what a model does not do as a reason for dismissing its results. Any model is necessarily an abstraction - in its ideal, leaving aside aspects of the richer empirical context that it considers to be insufficiently consequential for the operation of the particular causal mechanism that it aims to clarify. To avoid question-begging, a theoretical critique of the model must, thus, make a case that the omitted details are, in fact, consequential - either because their omission biases the results in some appreciable way or because it gives rise to some variety of internal incoherence (Laudan 1978). While it is often valuable to point out what is being omitted, a rejection of the model in response to such omissions but without the case that they are consequential is unwarranted. If the approach to analyzing deliberation as environment is to be a key element of deliberative democratic theory, it is also something that deliberative democratic theorists cannot afford.

Strategic Deliberative Behavior

It is, perhaps, unsurprising that the critiques of the game-theoretic conclusions that focus on strategic *deliberative behavior* are somewhat more philosophic in nature. As a class, these

critiques argue that the behavior characterized in the strategic models is inconsistent with the principal sources of participants' motivations, which tend to fall outside those models, and that such motivations trump the motivations explored by game theorists.

One such argument anticipates that the process of deliberation may lead participants to, effectively, "turn off" or set aside their self-interested motives in favor of reasoning with respect to a parallel set of desiderata - those that are more appropriate to the deliberative interaction. "Seeing that certain of my antecedent preferences and interests cannot be expressed in the form of acceptable reasons may help to limit the force of such preferences as political motives" (Cohen 1998, p. 199). This may happen for two reasons: either because we should be normatively committed to the view that "political justification requires finding reasons acceptable to others, understood as free and equal, who endorse that commitment" (Cohen 1998, p. 200; see also Scanlon 1998, Ch. 5), or because of what Elster (1995) calls "the civilizing force of hypocrisy" - we may simply be forced into making less self-serving arguments by the fact that such arguments do not take us very far in arguing with others, and over time internalize the value of other-regarding reason-giving as a way of coping with dissonance.

Elster's hypothesis would, likely, find some psychological support. However, there are good reasons to resist building a normative theory of deliberation upon this premise: it is equally compatible with conformity and raises questions about the compatibility of democratic procedures and individual autonomy (Johnson 1998, p. 172). A normative commitment to "honest" or "fully-revealing" deliberation does not suffer from these problems, but in a complex environment with private moral values and private veridicality, it must presuppose not simply abnegation of self-interest, but an extraordinarily high degree of abnegation of moral instrumental behavior as well. In both of our Jack and Jill examples, Jack's motivation is not self-interested: he is dedicated to a particular set of values and policies because

he believes them to be best on the merits. To the extent that, mindful of non-common values or private veridicality, Jack believes that deliberation will not lead to a consensus, in which either he and/or his interlocutors will necessarily be convinced by the other, a requirement that he always fully and honestly reveal may be expected to run into conflict with his commitment to the values and policies that he believes are most defensible.[17] It is that extremely demanding and quite controversial requirement that is, in effect, urged by this objection to the game-theoretic arguments. Embracing this requirement as a default seems premature both with respect to the status of moral-theoretic debate on its validity and as an empirical description of individual motivations.

Another argument targeting the deliberative behavior characterized in game-theoretic models concerns the legitimacy of the post-deliberative outcome. Its claim is that we want such outcomes to be perceived as legitimate, and that requires that participants fully and honestly reveal the arguments and information available to them (Habermas 1990). This claim is, in effect, comparing two states of the world: in state 1, the speaker, say, Jack, fails to speak or fails to reveal all the information available to him, and the outcome is less legitimate; in state 2, Jack reveals more information, and the outcome is more legitimate. To make sense of this argument, we must suppose, further, that, save for the effect on political legitimacy, Jack would prefer state 1 to state 2 - if it is not the case, then the invocation of political legitimacy is moot. It should be clear now that, in order for the argument to have bite, it must, in effect, endorse the following two claims: (1) interest in the perception of the outcome as legitimate overrides whatever interests Jack may have from revealing less information; and (2) perception of the outcome as legitimate is responsive to

[17]The epistemic issues in such problems are far from trivial. See Landa (2006) for a discussion.

how much information Jack reveals. Both of these claims strike us as highly controversial and contingent, at best. The first claim runs into a problem similar to the one we raised in connection with the previous argument: it presupposes a considerable constraint on Jack's commitment to the values or policies he believes to be most defensible. The second claim assumes, in effect, that it is known when Jack has relevant information and refuses to share it - an unlikely event in the context of our first Jack and Jill example, and certainly unreasonable as a default supposition.[18]

The last critical argument we consider proceeds on the premise that, in the same way that telling a lie is parasitic on telling the truth, strategic action is parasitic on "communicative action" - action that is oriented toward reaching mutual understanding in the course of deliberation (Habermas 1984). Because the strategic action is parasitic, the incentive-based behavior that game theorists are focusing on is, in essence, of secondary importance; it cannot, *ipso facto*, be the core of the deliberative practice (Heath 2001). To evaluate this objection, suppose, for the sake of argument, that strategic action is indeed parasitic, and consider the inference that that action is, therefore, of secondary importance. To the extent that this means that the strategic analysis of the incentives to reveal information is also of secondary importance, this argument says, in effect, that it does not matter how or why

[18]There is another variation on this argument (see Fearon 1998, p. 62 for a brief discussion): we want political outcomes to be perceived as legitimate, and legitimacy is enhanced by having losers in a vote "know exactly what reasons and arguments the winners had judged to be stronger in deciding the merits of the case." Here, there are two implied counterfactual states as well. In state 1, information is not revealed, Jack is better off, but losers don't know what it is that they lost to; in state 2, information is revelead, Jack is worse off, but losers know what it is that they lost to. To the extent that Jack has something to hide, it seems to us that the outcome in state 2 would be perceived as less, not more, legitimate.

speech is sincere when it happens to be such. As a theory of rational action, classical game theory is committed to the view that those how and why questions have answers that are not random, and that they relate systematically to the incentives faced by the participants; even when sincerity is unreflected, it is typically a consequence of the fact that the underlying environment is such that the incentives to misrepresent are not prominent.[19] In effect, the impulse behind the game-theoretic analysis of deliberation is to "earn" the sincerity by reconstructing it as equilibrium behavior rather than assuming it by default. As we argued above, the value of doing so is not only explanatory. Unless we understand the conditions under which the incentives in deliberative environments encourage agents to be sincere or fully revealing, as opposed to insincere or withholding of information, we cannot hope to offer a coherent (stable) normative argument for institutional design. If one of the goals of normative theories of deliberation is to offer such arguments, the game-theoretic analysis of incentives in deliberative environments is a useful tool that should be of primary importance even if strategic action is parasitic in the above sense.

Conclusion

Game-theoretic analysis demonstrates that deliberative democracy is not immune from incentive problems. Specifically, incentives to deceive or at least to refrain from sharing relevant information surface in many analytically distinct types of settings. What consequences should these observations be taken to have for the further development of deliberative democratic theory?

One possibility is to ignore the problems, assuming that it is reasonable to study only

[19] See Rubinstein (1991) and Landa (2006) for arguments about the relationship between perceptions of incentives, individual rationality, and internalized norms.

deliberation that is sufficiently idealized, that participants are unwilling to misrepresent or hide information. To the extent that expectations of sincere speech are critical for the deliberation-based theory of political legitimacy, this escape seems self-defeating, limiting the applicability of normative claims about deliberation to the most trivial of collective choice problems.

A second option is to accept the possibility of incentive problems but proceed on the assertion that these problems are only consequential in a narrow range of settings. As we have emphasized, however, the incentive problems surface in non cheap-talk settings, settings with repeated play, and settings with multiple senders and/or receivers. This option, then, also seems unwarranted.

A third response is to concede that these problems can be important, caveat one's arguments and conclusions so as to assume that incentive problems have been addressed (or will be addressed by someone else) and continue the research agenda without modification. This course of action is, in our opinion, ill-advised. The study of incentives leads to insights about how the particulars of a setting influence behavior. At present, the particulars of the deliberative environment tend to be absent from normative accounts. Unless we are convinced that these details are irrelevant, the current literature is likely to suffer from the pathology of formulating one-size fits all prescriptions. Based on the extant game-theoretic treatments, there is reason to believe that the details can matter, and ignoring the study of incentives means missing an opportunity to develop richer, more nuanced theories of deliberative democracy. A more careful approach that, in the way we characterized above, draws on the strengths of the normative analysis to elucidate axioms and ground rankings of tradeoffs, and on the precision of the game-theoretic framework to trace out how behavioral and institutional pieces fit together, appears most poised to take advantage of it.

References

[1] Austen-Smith, David. 1990. "Information Transmission in Debate." *American Journal of Political Science* 34 (1): 124-152.

[2] Austen-Smith, David, and Timothy Feddersen. 2005a. "Deliberation and Voting Rules" in David Austen-Smith and John Duggan, eds. *Social Choices and Strategic Decisions: Essays in Honor of Jeffrey Banks*. Springer.

[3] Austen-Smith, David, and Timothy Feddersen. 2005b. "The Inferiority of Deliberation under Unanimity Rule." *Northwestern University Typescript.*

[4] Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98 (4): 687-99.

[5] Bohman, James and William Rehg, eds. 1997. *Deliberative Democracy: Essays on Reason and Politics.* Cambridge: MIT Press.

[6] Bohman, James and William Rehg. 1996. "Discourse and Democracy: The Formal and Informal Bases of Democratic Legitimacy. *The Journal of Political Philosophy* 4(1): 79-99.

[7] Calvert, Randall L. 2006. "Deliberation as Coordination Through Cheap-Talk." *Washington University Typescript.*

[8] Cohen, Joshua. 1998. "Democracy and Liberty." In Jon Elster, ed., *Deliberative Democracy.* New York: Cambridge University Press.

[9] Cohen, Joshua. 1997. "Deliberation and Democratic Legitimacy." In J. Bohman and W. Rehg, eds., *Deliberative Democracy: Essays on Reason and Politics.* Cambridge: MIT Press, 67-92.

[10] Crawford, Vincent and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50: 1431-1452.

[11] Dryzek, John S. 2000. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations.* Oxford: Oxford University Press.

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

[13] Elster, Jon. 1995. "Strategic Uses of Argument." In Kenneth Arrow, ed., *Barriers to Conflict Resolution.* New York: W. W. Norton.

[14] Estlund, David M. 1997. "Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority." In J. Bohman and W. Rehg, eds., *Deliberative Democracy*, 173-204.

[15] Fearon, James. 1998. "Deliberation as Discussion." In Jon Elster, ed., *Deliberative Democracy.* Cambridge: Cambridge University Press, 44-68.

[16] Fishkin, James S. 1991. *Democracy and Deliberation: New Directions for Democratic Reform.* New Haven: Yale University Press.

[17] Gerardi, Dino, and Leeat Yariv. 2003. "Putting Your Ballot Where Your Mouth Is: An Analysis of Collective Choice with Communication." *Yale University Typescript.*

[18] Glazer, Jacob and Ariel Rubinstein. 2001. "Debates and Decisions: on a Rationale for Argumentation Rules." *Games and Economic Behavior* 36: 158-73.

[19] Guttman, Amy, and Dennis Thompson. 1996. *Democracy and Disagreement.* Cambridge: Belknap, Harvard Press.

[20] Habermas, Jürgen. 1996. *Between Facts and Norms.* Cambridge: MIT Press.

[21] Habermas, Jürgen. 1990. "Discourse Ethics: Notes on the Program of Philosophical Justification." In Habermas, *Moral Consciousness and Communicative Action.* Cambridge: MIT Press.

[22] Habermas, Jürgen. 1984. *The Theory of Communicative Action, vol. 1.* Boston: Beacon Press.

[23] Hafer, Catherine and Dimitri Landa. 2006. "Rules of Debate." *New York University Typescript.*

[24] Hafer, Catherine and Dimitri Landa. 2005. "Deliberation and Social Polarization." *New York University Typescript.*

[25] Hafer, Catherine and Dimitri Landa. 2003. "Deliberation as Self-discovery." *Journal of Theoretical Politics,* Forthcoming.

[26] Hardin, Russell. 1988. *Morality Within the Limits of Reason.* Chicago: University of Chicago Press.

[27] Heath, Joseph. 2001. *Communicative Action and Rational Choice.* Cambridge: MIT Press.

[28] Johnson, James. 1993. "Is Talk Really Cheap? Prompting Conversation Between Critical Theory and Rational Choice." *American Political Science Review* 87: 74-86.

[29] Johnson, James. 1998. "Arguing for Deliberation: Some Skeptical Considerations." In Jon Elster, ed., *Deliberative Democracy.* New York: Cambridge University Press.

[30] Knight, Jack and James Johnson. 2005. "On the Priority of Democracy: A Pragmatist Approach to Political-Economic Institutions and the Burden of Justification." *University Rochester Typescript.*

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

[32] Landa, Dimitri. 2006. "The Epistemic Theory of Toleration." In Ingrid Crepell, Russel Hardin, and Stephen Macedo, eds., *Toleration on Trial.* Lexington Books.

[33] Landa, Dimitri. 2006. "Rational Choices as Social Norms." *Journal of Theoretical Politics* 16 (4).

[34] Landa, Dimitri. 2005. "The Moral Economy of Deliberative Participation." *New York University Typescript.*

[35] Levy, Gilat. 2006. "Decision-Making in Committees: Transparency, Reputation and Voting Rules." London School of Economics Mimeo.

[36] Lipman, Bart, and Duane Seppi. 1995. "Robust Inference in Communication Games with Partial Proveability." *Journal of Economic Theory* 66: 370-405.

[37] Laudan, Larry R. 1978. *Progress and Its Problems: Towards a Theory of Scientific Growth.* Berkeley: University of California Press.

[38] Macedo, Stephen, ed. 1999. *Deliberative Politics: Essays on Democracy and Disagreement.* Oxford: Oxford University Press.

[39] Mackie, Gerald. 1998. In Jon Elster, ed., *Deliberative Democracy.* New York: Cambridge University Press.

[40] Manin, Bernard. 1987. "On Legitimacy and Political Deliberation." *Political Theory* 15 (3): 338-68.

[41] Meirowitz, Adam. 2005. "Designing Institutions to Aggregate Preferences and Information." *Quarterly Journal of Political Science,* Forthcoming.

[42] Meirowitz, Adam. 2003. "In Defense of Exclusionary Deliberation: Communication and Voting with Private Beliefs and Values." *Journal of Theoretical Politics,* Forthcoming.

[43] Mendelberg, Tali. 2002. "The Deliberative Citizen: Theory and Evidence." *Political Decision Making, Deliberation and Participation* 6:151-93.

[44] Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109 (2):231-265.

[45] Parfit, Derek. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

[46] Posner, Richard. 2004. "Smooth Sailing." *Legal Affairs*, January-February.

[47] Rubinstein, Ariel. 1991. "Comments on the Interpretation of Game Theory." *Econometrica* 59 (4): 909-924.

[48] Scanlon, T. M. 1998. *What we Owe to Each Other.* Cambridge: Harvard University Press.

[49] Stasavage, David. 2006. "Polarization and Publicity: Rethinking the Benefits of Deliberative Democracy." *Journal of Politics*, forthcoming.