Journal of Theoretical Politics

Deliberation as Self-Discovery and Institutions for Political Speech

Catherine Hafer and Dimitri Landa

The online version of this article can be found at:

Published by:

Additional services and information for *Journal of Theoretical Politics* can be found at:

Email Alerts: http://jtp.sagepub.com/cgi/alerts

Subscriptions: http://jtp.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 13 articles hosted on the
SAGE Journals Online and HighWire Press platforms):
http://jtp.sagepub.com/cgi/content/refs/19/3/329

DELIBERATION AS SELF-DISCOVERY AND INSTITUTIONS FOR POLITICAL SPEECH

Catherine Hafer and Dimitri Landa

ABSTRACT

We present a game-theoretic model of the social dynamics of belief change in which the (relevant) logically non-omniscient audience becomes convinced that the speakers' messages are 'true' because its own prior beliefs logically entail them, rather than – as in cheap-talk models – because the speaker is (endogenously) trustworthy. We characterize the equilibria of the game and consider how their aggregate informational properties change with the variation in the institutions determining the ability of the speakers to reach their audience. We find that for plausible restrictions on the distribution of arguments and on the corresponding policy preferences in society, the informationally optimal institutions are first-best implementable, inegalitarian with respect to the resource allocation across speakers, and assign priority to the (more) extreme argument- and policy-holders.

KEY WORDS ● deliberation ● non-Bayesian learning ● institutions

Introduction

One of the central claims of deliberative democratic theory is that deliberation – an unforced exchange of reasons aimed at the development of the most justified policy position or judgment – is a critical element of democracy. Provided that our judgments are responsive to the morally and politically relevant information communicated in the course of providing reasons for interlocutors' positions, individual and collective decisions untested by such exchanges are likely to be deficient in the eyes of some, if not all, citizens. If so, then these decisions – and, by extension, the institutional mechanisms generating them – may, in the absence of deliberation, fail to gain the requisite degree of political legitimacy (Manin, 1987; Habermas, 1996; Cohen, 1997).

To the extent that the quality of political decisions and the legitimacy of political practices are fundamental concerns of both political philosophy and

political science, an account of deliberation and of its determinants deserves a central place on the agenda of both. But while deliberation has, indeed, become one of the most actively explored issues in philosophy (Gutmann and Thompson, 1996; Bohman and Rehg, 1997; Elster, 1998; Macedo, 2000), it has received relatively little attention in rational choice social science. Not surprisingly, the extensive conceptual argumentation for deliberation has been matched with little headway in analyzing the optimal implementable institutional mechanisms that would substantiate in non-ideal political environments the properties claimed on its behalf (Johnson, 1998; Hardin, 2000).

Our aim is to contribute to closing the gap between these analyses of deliberation in two ways: by proposing an analytical model of deliberative practice that captures some of the key features of empirical deliberations falling outside the existing rational choice models, and by analyzing the equilibrium properties of a particular class of institutions in its context.

Arguments and Disagreement

The standard rational choice model of verbal communication – cheap-talk signaling – examines the extent to which agents can successfully communicate privately known empirical facts that are relevant to some collective decision. Versions of the cheap-talk model fix primitive preferences over outcomes and address the impact of communication on induced preferences over instrumental choices. Individuals may exchange messages about the content of their private information and, because that information is private, they can, within limits, engage in misrepresentation of their information. The determination of the extent or quality of deliberation in such models, then, turns largely on the credibility of the sent messages, which is induced by the proximity of the sender's primitive preferences to those of the individual receiver (Crawford and Sobel, 1982; Austen-Smith, 1990) or, in the case of a group of receivers, to those of the expected majorities within the group (Meirowitz, 2007), and by the sender's pivotalness both as a source of information and as a voter (Austen-Smith and Feddersen, 2002).

When the information transmitted in deliberation consists of privately known empirical facts relevant to the application of particular moral principles (self-interest, fairness, etc.), cheap-talk models capture important features of the deliberative interaction. However, as the following examples drawn from recent public discourse illustrate, deliberation commonly takes a form that cannot plausibly be construed as an exchange of previously unknown empirical evidence:

(1) *Although statistical justifications of racial profiling are not racist, they are wrong because they ignore the indirect costs of a practice that relies on racial selectivity:*

   Statistics abundantly confirm that African Americans – and particularly young black men – commit a dramatically disproportionate share of street crime in the United States. This is a sociological fact, not a figment of the media's

(or the police's) racist imagination ... [Nonetheless] politicians ... must be willing to demand equal treatment before the law even under circumstances in which unequal treatment is plausibly defensible in the name of nonracist goals ... [The reason is] the recognition ... that the presence of a racial factor in governmental decisionmaking gives rise to the presumption that officials may be acting in violation of someone's civil rights. (Kennedy, 1999)

(2) *The consenting adults argument for the legal recognition of gay marriage is consistent with the legal recognition of incest, and so cannot be a sufficient policy principle:*

If it's just about two consenting adults who love each other, why would we then deny a father and a daughter who's of age to get married? ... Why do we ban such things? It's because it's not good for society. (Wood, 2003)

(3) *Although the critics of the legal recognition of gay marriage are correct about the status of the consenting adults argument with respect to incest, it need not follow that the consequence is, therefore, necessarily socially deficient:*

Santorum and Scalia and the rest are absolutely correct that there is no principle which would deny two brothers from marrying that can allow two unrelated gay men from marrying. There is only taboo and prejudice .... All of the arguments about deformed children and abusive parents are red-herrings since the law could certainly protect against those problems while still allowing for gay same-sex siblings to do whatever they want. (Goldberg, 2003)[1]

Although these arguments include references to empirical facts, states of the world, etc., the deliberation they connote is, in effect, an exercise in 'experiential self-discovery'. When these arguments convince, they do so because they invoke as antecedents knowledge that hearers already have or that can be made self-evident to them as its deductive or inductive extensions.[2] In such deliberation,

---

1. Other, somewhat more esoteric, examples include most moral philosophical arguments, including John Rawls' (1978) for the uniquely fair character of his principles of justice. The example of Rawls' argument is particularly interesting in the context of our article if one interprets the device of the original position as de-biasing the perspectives of individual citizens, whose social and economic engagements make it difficult for them to see their way to the correct principles of justice (Hurley, 2000). On this account, as in our model, the normative argument is addressed to those whose ability to draw implications from their commitments (in Rawls' case, to fairness) is suitably impaired.

2. Calvert and Johnson (1998: 6) emphasize the prominence of non-cheap-talk arguments in motivating their view of deliberation as a mechanism of coordinating expectations in the context of multiple distributionally non-equivalent equilibria. We believe that deliberation quite likely serves that function but, in our view, it does so through preference transformation – effectively re-defining the initial game (though not the game-form). Significantly, our model of this process meets Calvert and Johnson's criticisms of preference-transformation views of deliberation: we assume neither the universal post-deliberative consensus (as in most primitive preference transformation theories) nor that cheap-talk communication of private information exhausts the deliberative practice.

the goal of the speaker is not to convince listeners that she and, by extension, her speech (argument/reason) are credible, but that propositions already known to the audience to be true are appropriately relevant to the decision at hand. Her arguments may fail to convince but, unlike in the case of the speaker with truly private information, largely not because the listeners are guarding against plausible misrepresentation.[3]

To underscore the distinction between the expert testimony-like nature of deliberation approximated by a cheap-talk model and deliberation exemplified by these arguments, we refer to the latter as '*deliberation as self-discovery*'. This kind of deliberation poses a substantial theoretical problem for the set of methodological assumptions employed in most rational choice models. If individuals are presumed to comprehend fully the implications of their knowledge, beliefs and preferences, how can the reiteration of some part of this information alter their policy judgments? Given that the assumption of logical omniscience obviates the need for arguments that are, from a logical standpoint, self-evident, developing an internally consistent explanation of that phenomenon necessitates a different informational model of agents' decision making.

We introduce such a model in Section 2. It differs from the standard models in two respects. First, it includes inherently persuasive messages, of the sort exchanged in deliberation as self-discovery. Second, it relaxes the usual assumptions of agents' logical omniscience in a way that is analytically precise and that corresponds to well-known experimental findings in social psychology. Taken together, these two features allow us to specify a model of deliberation that both retains the core elements of rational agency and allows us to analyze the analytically consistent possibility of deliberation as self-discovery within a broadly rational-actor framework.

We pursue this analysis in a relatively simple strategic model of group deliberation considered in Section 3. Its ontology of deliberation has two key elements that rationalize the possibility of 'self-discovery'. First, whatever its outward forms, deliberation as self-discovery is an attempt to make judgments consistent with policy-relevant commitments. As such, it is more than the exchange of messages in an institutionally suitable environment: individuals could be conversing without deliberating, and consequently, from a causal standpoint, deliberation proper must be thought of as an activity with a distinct set of (directly unobservable) determinants. This means, in particular, that the kind of learning that may be desired from deliberation should not be expected from the fact of conversation alone.

Second, we think of deliberation as entailing a trade-off between advocating a policy position by making arguments in support of it (a posture we refer to as

---

3. Our model is, however, compatible with the view that the willingness of the listener to entertain the speaker's argument for a particular position (as distinct from the willingness to accept the argument, once it has been entertained) depends on the speaker's credibility, itself induced by the listener's prior knowledge of the speaker's type.

'speaking') and contemplating the arguments presented by others (which we refer to as 'listening'). This division is a very common feature of large- and small-scale deliberative processes in which the degree of attachment to a given policy or argument and the knowledge of the corresponding issue domain often determine the extent to which individuals are willing to embrace the role of active 'carriers of the message' in that domain. In the model below, we capture this choice by making the cost of contemplating the arguments one hears and, possibly, changing one's induced preferences over policy as a consequence, the forgone opportunity to influence other agents' policy preferences in order to implement one's own ex ante policy judgment. The presence of such trade-offs is consistent with the variation in the cognitive status of attention and memory that is emphasized in psychological studies of the determinants of persuasion (Zaller, 1992; Lupia, 2002). It also explains why more or less rational individuals may, depending on circumstances, choose not to engage in costly information processing and, in conjunction with agents' limited cognitive abilities, it captures a plausible causal explanation for the phenomenon of deliberation as self-discovery.

Deliberative Institutions

The strategic context in the group deliberation game introduced in Section 3 provides a natural platform for the normative analysis of implementable deliberative institutions. Because disagreement over the best policy is carried through to agents' choices regarding deliberative participation, these choices need not lead to the best possible *aggregate* informational quality of post-deliberative decisions. We pursue the analysis of the issue of institutional choice framing such decisions in Section 4 of the article. We do so by considering a class of empirically plausible institutional arrangements that affect the informational quality of post-deliberative decisions by bolstering the reach of particular speakers.

As a heuristic abstraction, our model focuses on some of the features attributed to deliberative democracy without pursuing (though, arguably, being not inconsistent with) others – most notably, self-learning in the process of articulating an argument oneself. Correspondingly, our relatively modest goal is to clarify the precise purchase and implications of deliberation as self-discovery as a dialogical learning mechanism (Christiano, 1997; Fearon, 1998) without claiming to exhaust the causal landscape of deliberation in our analysis. In that spirit, our main results contain several significant, and perhaps surprising, implications for deliberative democratic theory. The first of these is the possibility (though not necessity) of multiple substantively distinct equilibria with different informational effects, and hence (in the informational causal model of deliberation) different policy outcomes and different normative properties. These different outcomes are possible even holding constant the institutional environment in which deliberation takes place. Second, we show that it is always possible to adjust institutional levers in order to select and implement uniquely the informationally best of these

outcomes for a given distribution of parameters. Third, the institutions supporting such optima are at variance with those that are typically advocated by deliberative democrats, both in that they are distinctly inegalitarian in their distribution of deliberative opportunities across parties publicly articulating different arguments, and in that they assign priority to the (more) extreme argument- and policy-holders.

Deliberation as Self-Discovery: The Informational Framework

In this section we briefly summarize the key features of our model of non-Bayesian deliberation, which is at the heart of the model of group deliberation developed in the following section. To provide a meaningful role for deliberation, we assume that agents are uncertain about the true merits of their alternatives. In particular, suppose that agent i is uncertain about her (true) ideal point $\hat{\pi}_i$ in the corresponding policy space. In order to operationalize that uncertainty in a manner consistent with the notion of deliberation discussed above, suppose that there is a set of reasons, R, each member r, of which, is considered to be potentially relevant to the determination of the true best policy by some member of the society. To capture the fact that not all members of a society necessarily find the same reasons compelling, let an agent's type be the set of all reasons that she would consider relevant to the determination of policy, $\mathcal{R}_i \subseteq R$, and suppose that agents are heterogeneous in type. Let $\mathcal{R}$ be the set of all possible types, and let $\hat{\pi}_i = \hat{\pi}(\mathcal{R}_i)$ definitively associate a true ideal point in the policy space with an agent's type, i.e., the function $\hat{\pi}$ associates a scalar with an unordered set. Let a true-type profile ω be an ordered list of the true type $\mathcal{R}_i$ of every individual i in the society, and let Ω be the set of possible true-type profiles.

The set of reasons, $\mathcal{R}_i$, that defines agent i, is partitioned (and may be re-partitioned) into a set of active arguments $\mathcal{A}_i$ and a set of latent arguments $\mathcal{L}_i$, with initial partition $\mathcal{A}_i^0$, $\mathcal{L}_i^0$. Agent i knows the profile of active arguments $\mathcal{A}$ (including her own active argument $\mathcal{A}_i$), but she does not know her latent argument $\mathcal{L}_i$; this limitation on her knowledge is the source of her uncertainty over $\hat{\pi}_i$.

When referring to the arguments that articulates reasons R, we identify them by their distinct theses, or distinct 'argument labels', for example, 'the human fetus is a human being'. Since an argument typically consists of a series of premises, inferences and conclusions, including a summary argument thesis/label, knowing the label is not the same as knowing, or being persuaded, or even necessarily knowing that one would be persuaded, by the argument for which it stands (as is the case in this example). However, the identification of an argument with a given label often enables agents to assess the probability of encountering the corresponding arguments in their social interactions, as well as to determine whether a given thesis, and so the arguments supporting it, could, in principle, comport with other arguments they know to be true. To simplify notation, we refer to the reason, the argument articulating that reason, and its

label as $r \in R$, with the understanding that, in a general case of non-active arguments, agents may, prior to deliberation, know only a label, and not whether they are convinced by its corresponding argument. Whether a non-active argument r is one's latent reason or not depends on whether one would accept the set of premises the corresponding argument employs. Thus, when referring to $r \in R$ as being in i's set of latent reasons, we mean that there is a set of premises that i accepts that would, on examination, imply the validity of the thesis r.

Let a society be composed of a finite set of types $\mathcal{R}$ and assume that agents know $\mathcal{R}$ and share common beliefs about the probability of each type, $\Pr(\mathcal{R}_i)$ for all $\mathcal{R}_i \in \mathcal{R}$. From these common beliefs they derive common conditional probabilities $\Pr(\mathcal{R}_i | \mathcal{A}_i)$. No new information about the primitive probabilities $\Pr(\mathcal{R}_i)$ is available from playing the game. Although $\mathcal{R}_i$ cannot change, the partition of it can. Let m_i be the set of i's received messages, $m_i \in 2^R$. If agent i observes (i.e., receives a message consisting of) an argument that is in her type $\mathcal{R}_i$, it becomes (or stays) active; if she observes an argument that is not contained in $\mathcal{R}_i$, then she is not persuaded by the argument and no change in the partition of $\mathcal{R}_i$ occurs. This cognitive capacity is summarized formally as

$$\mathcal{A}_i = \mathcal{D}(\mathcal{A}_i^0, \mathcal{R}_i, m_i) = \mathcal{R}_i \cap (m_i \cup \mathcal{A}_i^0). \tag{1}$$

This model is agnostic with respect to the nature of the inference entailed in the recognition of one's own latent argument. It could include the deductive closures of statements that the corresponding agents consider to be true but mistakenly believe to be irrelevant to the decision-making in question. But the recognition of latent reasons could be used to model an inductive structure of belief updating as well.[4] In fact, the only restriction (1) implies on the nature of the connections between the reasons or between the reasons and policies is that $\mathcal{A}_i$ and $\mathcal{L}_i$ be consistent, i.e., that $\mathcal{A}_i \nRightarrow \neg \mathcal{L}_i$.[5] However, in the interests of coherence, one may (as we do in the model of group deliberation below) add the following caveat: agents cannot give convincing articulations of arguments that rely exclusively on reasons that could be in their latent sets. Given our interpretation of the difference between active and latent arguments, this assumption is quite intuitive and, to some extent, inevitable: if it does not hold, these differences essentially disappear, and with them the very point of

---

4. For a sketch of an explicitly inductive model of analogical reasoning, see Aragones et al. (2001).

5. This may be interpreted to mean that this model of updating rules out the possibility that hearing an argument would lead you to abandon another argument that you had previously thought to be true. However, the model (1) allows one to get arbitrarily close to that possibility. To see this, suppose that on the basis of the information available to i about A_i and the distribution over ω, i chooses $\pi_i(A_i)$ that is ε away from $\pi_i(A_i, L')$, where L' is such that $(A_i, L') \in \omega$. Then, hearing and recognizing as true L'', such that $L'' \Rightarrow \neg L'$, would lead i to abandon what is, in effect, a belief that L' is true.

deliberation.[6] Formally, we assume $\forall i \in N$, $\forall r \in R$, if $\Pr(r \in \mathcal{L}_i | \mathcal{A}_i) \neq 0$, then $s_{ir} = 0$, where s_{ir} is the probability that i makes argument r.

Our basic ontology of learning is, then, that of recognizing latent reasons – reasons that agents are endowed with and would be able to embrace as 'their own' after recognizing their fit with other held beliefs, but which are not actively available to them prior to deliberation either for developing the corresponding policy position or for attempting to influence others.

We interpret our behavioral assumption with respect to policy belief updating as saying that agents' policy positions are 'sticky' in the following sense: they change their policy positions only when they can give or understand a sound and valid propositional support for the newly adopted positions, at which point they switch to the policy implied by the conjunction of that (previously latent) reason and their initial active reason. The 'sound and valid propositional support' is a valid argument that proceeds from premises that are shared by the listener (or, at least, from the premises that are not contradicted by the listener's prior beliefs). The listener's 'latent' reason corresponds precisely to that argument, and, with the object of learning effectively entailed in her beliefs, her response embodies what is, arguably, the core element of any philosophically defensible account of rationality – 'as a norm, as a second-order disposition of the following kind: once one becomes aware that one has fallen into irrationality, one will tend to adjust one's beliefs, attitudes, and actions such as to make them more rational' (Føllesdal, 1982: 309; Landa, 2006). Thus, although our model of agency clearly departs from logical omniscience, it does so without abandoning the overall philosophical framework of rationality.

The belief 'stickiness' that constitutes this departure from logical omniscience makes our agents akin to Dr Watson, who, to Sherlock Holmes' repeated frustration, appears unable to infer the true alternative from the impossibility of its complements. To identify the right alternatives, they, like Watson, need direct evidence, $\mathcal{L}_i$. As we show elsewhere (Hafer and Landa, 2005), agents characterized by (1) systematically and exclusively fail the condition of *Negative Introspection* – they do not know what they do not know.[7] Upon hearing an argument that relies on a

---

6. It is, however, entirely possible that agents' thought processes prior to public deliberation, including their attempts to articulate to themselves the arguments they know to be true, leads them to inferences that increase the size of the set of active arguments. We do not model this process here, but it is consistent with everything that follows.

7. Let an agent's set of considered arguments μ_i consist of all the arguments that she has ever heard, including her own initial active argument, $\mu_i = m_i \cup A_i^0$. Let $\mu = (\mu_1, \ldots, \mu_i, \ldots, \mu_n)$ be the profile of sets of considered arguments and M be the set of all such profiles. Then the informational structure described above is representable as a possibility correspondence (Geanakoplos, 1989) $P_i : \Omega \times M \to \Omega \times M$, which associates with the true state the set of all states that i deems possible in that state. Letting $E \subseteq \Omega \times M$ be an event, we can, then, derive from our possibility correspondence its correlated knowledge function identifying the set of states in which every state that i deems possible is contained in E. We show that this knowledge function satisfies all of the epistemic logic axioms axiomatizing Bayesian updating, except Negative Introspection.

reason that is not in their 'latent' set, our agents would not (as true Bayesian agents would) infer from the path of play and their knowledge of the distribution of types that they are less likely to be a particular type that corresponds to the unconvincing argument and so must assign greater likelihood to being one of the complementary types. In effect, then, they will act as if they do not know that the information available to them has implications for what latent reasons to they must or are likely to agree with, and so will fail to update their beliefs about their induced ideal policies accordingly (like Dr Watson and unlike Holmes – the indisputable Bayesian hero – in the famous episode involving the dog that did not bark in the night).

One rationale for our assumption of violation of Negative Introspection is empirical plausibility. Long before cognitive and social psychologists gave us the language and the systematic evidence to describe this phenomenon, Conan Doyle had noticed that the embodiment of an educated and intelligent everyman systematically lacked Holmes' impressive Bayesian credentials. Modern psychologists identify the cognitive bias in favor of one's own currently held convictions – whereby, in order to trigger a change in one's prior position, the argument must leave little to ambiguity – as one of the most robust experimental findings (Tetlock, 1992; Zaller, 1992; Baron, 1994; Dawes, 1998; Rabin, 1998). As the authors of one of the seminal studies note, agents 'may even come to regard the ambiguities and conceptual flaws in the data opposing their hypotheses as somehow suggestive of the fundamental correctness of those hypotheses' (Lord et al., 1979, p. 2099). Studies of hypothesis testing (Wason, 1968, 1977; Baron, 1994: Ch. 13) find systematic evidence of a reluctance to see even the possibility of making valid inferences from disconfirmations of the consequent. Finally, in an experiment based on the model presented here, Dickson, Hafer, and Landa (2005) find that subjects' speaking and listening behavior conforms considerably more closely to the prediction for non-Negatively Introspective agents than to that for Bayesian agents.

Aside from empirical psychological plausibility, positing our agents to be like Dr Watson rather than like Holmes plays a central explanatory role, which we alluded to in the Introduction. The shortcomings of their rationality provide justification for the kind of deliberation modeled in this article, and are, in turn, implied by the need for a consistent model of agency in the analysis of deliberation as self-discovery. If agents update their beliefs only in response to direct arguments – i.e., arguments that draw out the logical inference from propositions they already hold true – then their failure to observe some such argument explains their need for deliberation as self-discovery. We model their updating as a result of such deliberation, then, precisely in the way that comports with this rationale.

A Model of Group Deliberation

The basic interaction of interest to us is as follows. First, on the basis of the information available to them about their own and other agents' ex ante preferred policies and about the arguments they could adduce to support them, agents

simultaneously choose how to distribute their scarce deliberative resources between 'speaking' and 'listening'. As a function of the combination of such choices by all agents and the institutional properties mediating them, agents receive varying degrees of exposure to the arguments supplied by their speaking counterparts; although all the speakers' messages are broadcast to all listeners, not all messages are necessarily heard by every listener. Next, they update their beliefs in a way described in the previous section, and, at the last stage of play, make the corresponding policy choices.[8]

The effectiveness of communication depends on the resources allocated to sending and to receiving messages (Zaller, 1992: Ch. 7), and we assume that each agent has a fixed amount of resources to expend in deliberation, normalized to 1. Let λ_i represent the amount that agent i allocates to receiving messages and $1 - \lambda_i$ the amount allocated to sending messages. In this way, we capture the costs entailed in deliberation as forgone deliberative opportunities.[9] Let N be the set of agents, $|N| = 4$. Let $s_i = (s_{i1}, \ldots, s_{i|R|})$ be a $|R|$-dimensional vector, where s_{ir} is the probability that i sends message r and the components of s_i sum to 1. Alternatively, s_{ir} may be interpreted as the proportion of $(1 - \lambda_i)$ committed to sending r. Let $s = (s_1, \ldots, s_4)$ be the profile of probability distributions over messages. The relationship between s and the profile of messages received m is mediated by the institutional structure K, which may be thought to comprise the rules that enable their speakers to make themselves heard (e.g., allocating time in front of the microphone, money to buy time, etc.), and the individual choices of λ. The probability that i hears a particular reason r is determined by s, K, and λ. Given this environment, agents have the information necessary to determine the probabilities of possible outcomes of deliberation.

In their decisions regarding the choice of π and λ, we assume agents to be expected utility maximizers, given their beliefs about their true reasons $\mathcal{R}_i$ (i.e., their beliefs about their $\mathcal{L}_i$, given their certain knowledge of $\mathcal{A}_i$) and thus about their ideal point $\hat{\pi}_i = \hat{\pi}(\mathcal{R}_i)$.[10] We assume a unidimensional policy space $\mathbb{R}$.

---

8. Although our model includes only one explicit round of messaging, it is compatible with the interpretation whereby the sender engages in a series of prompted elaborations of her message – culminating in the transmission of what may be thought of as the 'best possible articulation' of that argument.

9. Introducing leisure as a separate cost of a deliberative activity will have the effect of increasing the number of equilibria: in particular, for some ranges of parameters, there will exist equilibria in which some of the players choose leisure, as well as equilibria in which all of the players will choose leisure. However, all of the equilibria we characterize below will have analogues in which the consequences of deliberation are the same. Moreover, the tradeoffs between speaking and listening that we are characterizing in the article as well as the general conclusions about the relationship between informational efficiency and equality of deliberative opportunity will remain the same.

10. We thus bracket all other possible deviations from logical omniscience. Our strategy here is similar to that of Rabin and Schrag (1999: 38, 2), whose model incorporates one particular cognitive shortcoming, leaving rational behavior intact everywhere else in the model in order to isolate the effects of that shortcoming (or, in our case, of the social practice that corresponds to it).

Each agent cares primitively only about his and others' policy choices, $\pi = (\pi_1, \ldots, \pi_4)$, with $\pi_i \in \mathbb{R}$, and enjoys the utility $u_i(\hat{\pi}_i, \pi)$, which is a function of his true (full-information) policy preference, and the vector of actual policies choices made by members of the society, including i.

To obtain a more or less sharp equilibrium characterization, we must specify the shape of individual utilities and the relationship between their reasons and policy ideal points. (The conjunction of these restrictions is broadly equivalent to utility specification in the standard Bayesian models.) Rather than explicitly modeling preference aggregation, we assume that agents' utility is additively separable with respect to the policy choices of different members of the society. This assumption corresponds to what Baron (2003) refers to as 'private politics' – the use of the public or political sphere to affect the private choices of other citizens – and allows us to focus our analysis on deliberative behavior, setting aside the usual complications of constrained choice, including strategic voting, within specific institutions of preference aggregation.[11] The setting of private politics lends further credibility to the assumption of a unidimensional policy space since the private implementation of choices effectively prohibits credible bargaining over policy compromises across possible issue dimensions.

We assume, moreover, that choices by individuals other than i are weighted equally in i's utility and have, for i, no more weight than i's own choice. Formally, then,

$$u(\hat{\pi}_i, \pi_i, \pi_{-i}, ; \theta, n) = -(\hat{\pi}_i - \pi_i)^2 + \frac{\theta}{n-1} \sum_{j \in N \setminus i} (-(\hat{\pi}_i - \pi_j)^2) \qquad (2)$$

where $\theta \in [0, (n-1)]$. The quantity $-(\hat{\pi}_i - \pi_i)^2$ measures i's disutility from the distance between her chosen and true policies, and $\sum_{j \in N \setminus i} (-(\hat{\pi}_i - \pi_j)^2)$ captures i's disutility from the deviation of others' final policy choices from i's own true policy. The ratio $\frac{\theta}{n-1}$ could be thought of as a measure of how much i values other persons' choices relative to her own, with $\theta = n-1$ when i thinks of her own choice as having the same moral weight for her as that of each of the other members of the group, and $\theta = 0$ when i's moral system is effectively autarchic, i.e., when policy choices by others do not enter into i's welfare.

We consider a set of possible reasons with four elements, $R = \{A, B, C, D\}$, and a set of possible types consisting of three two-element subsets of R;

---

11. From the deliberative-democratic perspective, our representation of individual utilities is, arguably, most plausible in deliberations within 'civil society' or the 'public sphere' (Habermas, 1989). Examples of issues, deliberations on which would most immediately fall into the domain of private politics, include the desirability of obtaining abortions (holding constant the official policy); the workforce participation of women; the value of post-secondary education; choices concerning child-bearing, etc. More generally, our formulations of individual preferences could be thought of as the first-order approximation of preferences in the context of public politics, i.e., politics in which a single binding decision is made following deliberation.

$\mathcal{R} = \{\{A, B\}, \{B, C\}, \{C, D\}\}$. The initial partition of $\mathcal{R}_i$ into $\mathcal{A}_i^0$ and $\mathcal{L}_i^0$ is such that each agent has exactly one active and one latent argument. Unless specified otherwise, we let $\bigcup_{i=1}^4 \mathcal{A}_i^0 = \{A, B, C, D\}$, so that each possibly persuasive (to someone) argument has a proponent in the society, i.e., each of the four agents has a different active reason. Although this restriction is not without loss of generality, it has benefits beyond the increase in the technical tractability of the model. In particular, it allows us to avoid having to make contestable assumptions on the nature of (potentially strategic) coordination among agents with the same active argument on 'getting out the message', and focus our attention on the case of full argument representation, which is often implicitly assumed in political theory discussions. Since one of our main results argues against the equal institutional treatment of proponents of different reasons, this assumption is, effectively, biased against our findings.

We restrict our attention to a relationship between types and policy ideal points, i.e. a mapping of the set of types into the policy space, such that there is no argument that is shared by the extreme policy types but not by the moderate type (a property that we refer to as *connectedness*). This restriction does not constrain agents from sending messages across the argument spectrum – i.e., it does not require that i, such that $\mathcal{A}_i = A$, necessarily send the A message or target as her audience the agents whose (ex ante) policy is closest to her own. What connectedness does rule out are cases in which the opposite ideological extremes could be persuaded by the same argument on behalf of two different policies *even though the policy moderate finds that argument unpersuasive*. For example, a mapping of types into ideal policy points s.t. $\hat{\pi}(\{A, B\}) < \hat{\pi}(\{A, C\}) < \hat{\pi}(\{B, C\})$ is *not* connected because B is persuasive both to the leftist *and* to the rightist but *not* to the moderate. Though connectedness is a formal constraint, it seems empirically plausible for a wide range of cases of issue deliberation.[12]

Proceeding with these assumptions and eliminating redundant combinations of sets of types and ideal policy mappings, we arrive at the following two possibilities: $\hat{\pi}(\{A, B\}) < \hat{\pi}(\{A, C\}) < \hat{\pi}(\{A, D\})$ and $\hat{\pi}(\{A, B\}) < \hat{\pi}(\{B, C\}) < \hat{\pi}(\{C, D\})$.[13] A brief glance at the incentives induced in the deliberation game with the former indicates that the game is dominance-solvable. Although agents

---

12. This condition becomes more demanding as the cardinality of the set of true types increases. The plausibility of high-cardinality sets depends on one's view of the number of distinct internally consistent types in the population with respect to a given issue. Our guess is that, on most issues, that number is rather small.

13. All possible combinations of two out of four reasons is a set of six possible two-reason types, $\{AB, AC, AD, BC, BD, CD\}$, which generates 20 possible sets of three types, R, each of which gives rise to six possible permutations under $\hat{\pi}$. Out of the resulting 120 combinations of mapping $\hat{\pi}$ on domain R, only four are distinct up to symmetric transformations with respect to the reason names. In addition to the two possibilities discussed in the text, they include two unconnected mappings such that $\hat{\pi}(\{A, B\}) < \hat{\pi}(\{A, C\}) < \hat{\pi}(\{B, C\})$ and $\hat{\pi}(\{A, B\}) < \hat{\pi}(\{A, C\}) < \hat{\pi}(\{B, D\})$. The exposition is straightforward and omitted here.

whose active reasons are B, C, and D do not know reason A, the policy consistent with their active reason is unique, and so they know their correct policy choices with certainty. If so, then, agents whose active reason is A have no messages of interest to others to communicate. Correspondingly (using the subscript to denote the agents with the relevant initial active reasons) $\lambda_A^* = 1$, and $\lambda_B^* = \lambda_C^* = \lambda_D^* = 0$. This case is, therefore, uninteresting from a strategic standpoint and weakly independent of the institutional circumstances framing the individual choices.

The remainder of the article, therefore, assumes the latter case, with the society containing three possible types, $\mathcal{R} = \{\{A,B\}, \{B,C\}, \{C,D\}\}$, mapped into the space of policies by $\hat{\pi} : \mathcal{R} \to \{\hat{\pi}(\{A,B\}) = -1, \; \hat{\pi}(\{B,C\}) = 0, \; \hat{\pi}(\{C,D\}) = y\}$, where $y > 0$.

An agent may, on the basis of her evaluations of the likelihood that she is of a given true type $\mathcal{A}_i \cup \mathcal{L}_i$, choose π_i to be any point in $\mathbb{R}$; because her true ideal point is in the interval $[-1, y]$ with certainty, policies outside this interval are dominated, and so without loss of generality we can restrict our attention to $[-1, y]$. Agents have common beliefs, given $\mathcal{R}$, about the stochastic process generating $\mathcal{R}_i \in \mathcal{R}$. Let $p_{\mathcal{R}_i} := \Pr(\mathcal{R}_i)$, and $p = (p_{\mathcal{R}_1}, p_{\mathcal{R}_2}, p_{\mathcal{R}_3})$; the list of types is exhaustive, so $\sum_{\mathcal{R}_i \in \omega} p_{\mathcal{R}_i} = 1$. Given the distribution of true types and the available information about her own type, A_i, agent i can identify the policy that maximizes her expected utility. Given the vector of equilibrium policy choices in the post-deliberation continuation game, other agents' (simultaneous) deliberative behavioral strategies, and beliefs about the true-type profile (conditioned on the given active-reasons profile), each agent can choose the messages she sends and allocate her deliberative resources optimally.

Finally, $K = (K_A, K_B, K_C, K_D) \in [0,1]^4$ is our vector of institutional parameters (with each agent being notationally identified by her active argument). An increase in the institutional parameter for a given speaker increases the likelihood that that speaker's arguments are heard. We assume that the probability that any agent i will hear a particular reason r sent by an agent j is increasing in the resources that i dedicates to listening and in the resources that j dedicates to speaking. In particular, that probability is $(1 - \lambda_j)K_j\lambda_i$, and given the profile of probability distributions over sent messages s, the profile of individual allocations λ, and the institutional parameter K, the probability that i hears a particular argument r from any speaker is $\sum_{j \neq i} s_{jr}^* (1 - \lambda_j) K_j \lambda_i$.[14]

Equilibrium

We begin our analysis of the deliberation game by characterizing its equilibria. As in the usual Perfect Bayesian equilibrium concept, we require that each

---

14. We discuss the robustness of our results with respect to the shape of this communication technology in the text that follows.

agent's equilibrium behavioral strategy maximize her expected utility, given her beliefs and the strategies of other agents, at the point in the game at which she is called upon to act. (For the formal definition of the equilibrium behavioral strategy, see Proof of Proposition 2 in Appendix.) Contrary to standard Bayesian equilibrium concepts, however, our agents do not update their beliefs in a manner consistent with Bayes' Rule; rather, we assume that beliefs change in accordance with the cognitive model described in the previous section. We restrict attention to equilibria in undominated strategies. Let vector (λ^*, π^*, s^*) be a behavioral strategy profile that satisfies all the equilibrium requirements. Then:

PROPOSITION 1: *In the game described above:*

   (i) *i's equilibrium policy choice, as a function of $\mathcal{A}_i$, is $\pi_i^*(\{A\}) = \pi_i^*(\{A, B\}) = -1$, $\pi_i^*(\{D\}) = \pi_i^*(\{C, D\}) = y$, $\pi_i^*(\{B, C\}) = 0$, $\pi_i^*(\{B\}) = -\frac{p_{AB}}{p_{AB} + p_{BC}}$, and $\pi_i^*(\{C\}) = \frac{p_{CD}}{p_{CD} + p_{BC}} y$;*
   (ii) *for all $i \in N$, $s_{ir}^* = 1$ for $r \in A_i$ and $s_{ir}^* = 0$ for $r \notin A_i$;*
   (iii) *the equilibrium vectors λ^* are as depicted in Figure 1, with the values of $\tilde{\lambda}_B(p, K, y)$, $\tilde{\lambda}_C(p, K, y)$, $G(\theta, y, p)$, and $F(\theta, y, p)$ derived in the Appendix.*

Proof. See Appendix.

As Figure 1 shows, the group deliberation game has different unique equilibria on three subsets of parameter values and multiple (three) equilibria on the remaining subset. Given the form of agents' utility functions, their (post-deliberative) policy choices in all of these equilibria are, not surprisingly, dominant strategies. More interesting is the fact that, for all parameter values, all agents send messages consisting of their active reasons. Given our model of updating, doing so can only move the listeners toward the respective speakers, whereas sending other messages would, in expectation, have the opposite effect. 'Sincere' speech thus emerges in our model not as a consequence of myopic speakers, but of myopic listeners.[15] Another feature common to all equilibria is that agents with extreme reasons (A or D) never listen in deliberation. Although such agents do not know their latent reasons (in the sense of being able to articulate them), they can infer with certainty the policies that correspond to their active reasons, and so cannot learn any piece of information that would make them change their policy choices. Being convinced by hearing their latent reasons has, thus, no impact on their utility, and they are always willing to forgo hearing them for the chance of convincing a moderate to move toward them.

The only difference in equilibrium play is, then, in how much listening and speaking is done by the pre-deliberative policy moderates. For some conditions on the primitive features of the deliberative environment, one or both types of
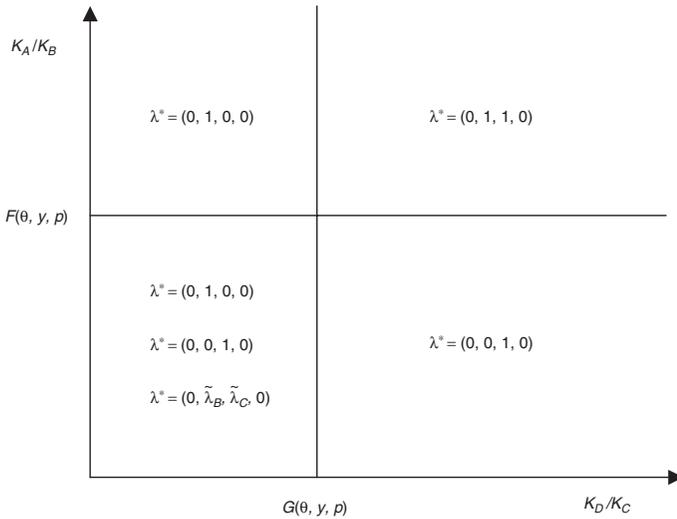
---

15. Compare to Mackie (1998).

Figure 1. Equilibrium Vectors λ^* of the Group Deliberation Game

moderate will receive sufficient benefits from listening to an extreme argument that they will strictly prefer listening to speaking regardless of the other moderate's behavior. For other conditions, a moderate will prefer to listen only if the other moderate speaks enough (in addition to the relevant extremist's speaking), and will otherwise prefer to speak; when that is true for both moderate types, multiple combinations of behaviors can be self-enforcing.[16] The simultaneous existence of three equilibria in the lower left quadrant of Figure 1 raises the question of equilibrium selection. In the next section, we address this question from a normative standpoint but first consider the policy consequences of deliberation.

To get a measure of policy change, we compute the aggregate expected deliberation-induced policy shift by summing the differences between pre- and expected post-deliberative policy positions. Without loss of generality, let $E[\Delta Right]$ be the shift rightward, so that

$$E[\Delta Right] = E\left[\sum_{i \in N}(\pi_i^*(\mathcal{D}(\mathcal{A}_i^0, \mathcal{R}_i, m_i)) - \pi_i^*(\mathcal{A}_i^0))\right]$$

---

16. The possibility of multiple equilibria is preserved for any communication technology in which both listening and speaking have diminishing marginal returns, i.e., such that $\Pr(r_j \in m_i)$ is weakly increasing and concave with respect to λ_i and $\Pr(r_i \in m_j)$ is weakly decreasing and concave with respect to λ_i. The specification of $F(\theta, y, p)$ and $G(\theta, y, p)$ changes, but the substantive conclusions are the same. The details are available from the authors upon request.

Note that this measure produces the value of the *net* aggregate shift, so that right-ward and leftward changes of the same magnitude produce a net aggregate shift of 0. The following proposition summarizes our findings on the manifest ideological effects of deliberation, as measured by $E[\Delta Right]$.[17] Because equilibria are distinct in λ^*, we henceforth identify them by λ^*.

PROPOSITION 2:

(i) *For every equilibrium, the expected aggregate post-deliberative rightward shift in policy is non-increasing in K_A and K_B and non-decreasing in K_C and K_D;*

(ii) *Both the direction and magnitude of the aggregate shift in policy, as well as the response of each to changes in y, vary with the equilibrium being played.*

Proof. See Appendix.

As part (i) of the proposition shows, greater control over deliberative resources has, not surprisingly, a favorable effect on the direction of the post-deliberative policy drift across the equilibria. What is less expected, perhaps, is the effect of the interaction between the equilibrium and y on the direction and magnitude of aggregate ideological shift. In particular, in the appendix, we show that in $\lambda^* = (0, 1, 0, 0)$ and $\lambda^* = (0, 0, 1, 0)$, the direction of the aggregate shift is independent of y. In $\lambda^* = (0, 1, 1, 0)$, the aggregate shift is rightward if and only if y is sufficiently large. Moreover, in $\lambda^* = (0, 1, 0, 0)$, the magnitude of the aggregate shift is independent of y; in $\lambda^* = (0, 0, 1, 0)$, the magnitude of the shift is increasing in y, regardless of the direction of the shift; in $\lambda^* = (0, 1, 1, 0)$, the magnitude of the shift is decreasing in y when the aggregate shift is leftward but increasing in y when it is rightward.

The intuition for these differences is as follows. Because in the equilibrium $(0, 1, 1, 0)$, agent B is exclusively listening, C is not receiving any B messages and so can only move to the right. The more radical the right-most position (i.e., the greater the y), the greater will be the right-ward tendency. By contrast, in the equilibrium $(0, 0, 1, 0)$, C is receiving B messages and so can move to the left. More-over, the greater y, the greater is the potential move to $\pi = 0$, since rational expectation of high y are, all else equal, inducing a more right-leaning pre-deliberative policy position. The consequence is a somewhat subtle conclusion: the more extreme the right, the greater is the expected post-deliberative move to the left.

The intuition for the invariance of the ideological shift with respect to the variation in y in the equilibrium $(0, 1, 0, 0)$ is in the fact that, as Proposition 1

---

17. We confine our analysis of the impact of changes in primitives on aggregate ideological outcomes to points at which these changes affect agent behavior (as captured by the best-response correspondence) smoothly, and so do not derive comparative statics for the aggregate outcomes associated with $\lambda^* = (0, \tilde{\lambda}_B, \tilde{\lambda}_C, 0)$.

shows, that equilibrium exists only if the right policy extreme is sufficiently moderate (i.e., y is sufficiently low). When that condition holds, it is rational for the right-moderate agent C to forgo listening entirely, and so the local variation in y has no effect. However, a substantial change in y away from the range supporting this equilibrium would lead to an equilibrium switch and a corresponding change in the expected aggregate policy shift.

Institutional Choice and Informational Welfare

Suppose that a government (or a constitutional mechanism that entails an effective instrument for enforcing the government's actions) is choosing a policy that determines K – in effect, the background institutional 'rules' in the context of which the deliberation game is played. A particularly appealing candidate for such rules is the maximization of the aggregate informativeness of post-deliberative individual positions, or, equivalently, minimization of the aggregate deviations of agents' post-deliberative policy choices from their own true (fully informed) policy preferences. K chosen in this fashion would epitomize the familiar reasons for citizen sovereignty and respect for diversity of opinion, and against government paternalism and the non-neutral manipulation of information.[18] In the remainder of this section, we consider the features of institutions that meet this criterion. Formally, then, let aggregate expected error in policy choices $I = E[\sum_i (\hat{\pi}_i - \pi_i)^2]$ be our measure of the informational inefficiency of deliberation.

The I-minimization criterion of institutional design can be used as a normative criterion of equilibrium selection and as the criterion for the selection of institutions K that generate optimal implementable outcomes. In particular, given the existence of multiple equilibria, we may rank them in relation to our measure of informational welfare and, conditional on their equilibrium implementability, use the behavioral levers afforded by the choice of K to direct play toward the most informationally efficient equilibrium.

Our next proposition addresses the nature of the behavioral targets for our normative analysis and their implementability, i.e., asking what patterns of individual behavior yield the most informationally efficient outcomes, and how much equilibrium play ties the hands of the institutional designer. The proposition shows that for any given set of primitives, the most informationally efficient equilibrium corresponds to the interlocutors' corner allocation of λ (i.e., an allocation in which every individual strictly prefers only to speak or only to listen).

---

18. Significantly, this criterion is distinct from that of maximizing the extent deliberation. Since deliberation is a means rather than an ultimate end, and it has, as modeled here, a publicly appreciable opportunity cost, this distinction is not surprising.

PROPOSITION 3:

> (i) *One of the corner allocations λ is always more informationally efficient than the interior equilibrium;*
> (ii) *Each of the three corner equilibria is uniquely implemented under the values of K that minimize I, given the corresponding equilibrium λ*;*
> (iii) *The most informationally efficient allocation of resources can be uniquely implemented.*

Proof. See Appendix.

This proposition may be seen to yield two important implications for the analysis of deliberation. First, the (more) procedurally egalitarian or symmetric equilibrium, in which both parties that stand to learn from listening engage in some speaking and some listening, is always inferior, from the standpoint of informational efficiency, to one or another of the very asymmetric corner equilibria, in which only one of two parties that stand to learn from listening actually does listen and listens exclusively, while the other exclusively speaks. Significantly, this result is independent of the specifications of the underlying parameters of the model, though the identity of the most informationally efficient equilibrium is not. If the causal purchase of deliberation is informational, this proposition establishes the appeal of managing public deliberation through explicit *asymmetric* restrictions on the roles of deliberating parties.

The intuition for this result is as follows. B's choosing to allocate positive amounts of her deliberative resources to both activities implies that, given C's behavior, B's marginal expected return from attempting to persuade others (speaking) and contemplating the verity of others' reasons (listening) are equal.[19] But, while the marginal benefit of listening is the same whether evaluated according to the informational criterion $(\min E[I])$ or the individual choice criterion $(\max E[u_i])$, the marginal benefit of speaking is different. This discrepancy occurs for two reasons. First, the individual decision maker cares less about the choice of any other individual than she does about her own, i.e. $\theta < n - 1$, while the informational criterion values the choice of each individual equally. Second, the individual decision maker is evaluating the desirability of others' ultimate policy choices against her own ideal policy point, whereas the informational criterion is evaluating the desirability of each individual's choice against that individual's ideal policy point. Either of these factors is enough to ensure that if the individual is indifferent between speaking and listening, the marginal expected decrease in the aggregate error from one activity must be greater than that of the other. It

---

19. The indifference is induced by the linearity of $E[u(B)]$ in λ_B, which is plausible in this model. Doing away with the linearity would give Bs strong preference, but not alter the flavor of the result, since the divergence in the social and in the individual incentives would remain.

follows that one of the corner equilibria better satisfies the informational criterion than does the equilibrium in which all agents devote some resources to speaking.

By way of better understanding this result, it is worth underscoring its formal and substantive connection to our account of deliberation as self-discovery. As we show elsewhere, if our agents were Bayesian, we would see no speech in equilibrium.[20] The intuition is as follows. A Bayesian will update his beliefs in response to an unpersuasive reason as well as a persuasive one. Because players' pre-deliberative positions are in expectation optimal (that is, because they [correctly] incorporate their uncertainty about their true ideal points), and because the speaker obtains diminishing marginal utility from the proximity of the listener's choice to the speaker's own ideal point – a standard assumption in the spatial models of politics – the expected harm to the speaker associated with the listener resolving her uncertainty outweighs the expected benefit to the speaker of the listener doing so. Because no players will speak, no deliberation will happen, and so the issue of informational optimality of equilibria with Bayesian players is moot.

This said, it is the case that one can construct a traditional Bayesian cheap-talk model of information transmission that would yield a result that would be qualitatively similar to Proposition 3. However, given that model's approximation of what we referred to above as 'expert deliberation', such a result would be neither substantively surprising (after all, some of us are, indeed, experts and others may very well be less so) nor all that controversial to deliberative democrats. Because the model of deliberation as self-discovery more closely approximates what arguably happens in day-to-day deliberative interactions between citizens (interactions in which the speaker's access to private information is a relatively atypical defense of a proffered claim), the participants are best thought of as much more equal in terms of what they can offer to each other. In that context, a result that says that inequitable opportunity to speak is informationally efficient is, therefore, considerably more interesting.

Indeed, from the normative standpoint, this result is rather striking. In particular, it appears to be at odds with one of the strongest intuitions in the deliberative democracy literature. As one of its leading contributors put it:

> In ideal deliberation, parties are both formally and substantively equal. They are formally equal in that the rules regulating the procedure do not single out individuals. Everyone with the deliberative capacities has equal standing at each stage of the deliberative process. Each can put issues on the agenda, propose solutions, and offer reasons in support of or in criticism of proposals. And each has an equal voice in the decision. The participants are substantively equal in that the existing distribution of power and resources does not shape their chances to contribute to deliberation, nor does that distribution play an authoritative role in their deliberation. (Cohen, 1997: 74)

---

20. Hafer and Landa (2005). Although the specification of the state space is somewhat different, the same argument used to establish the result in that paper is valid in the present model as well.

The intuition behind this statement is both straightforward and attractive: true policy or judgment cannot be manufactured by anything other than reason, and denying equal public airing of claims to it seems to suggest either paternalism or efforts to manipulate the outcome (why else would the voice be suppressed if not because of the possibility that the audience may find it [dangerously] appealing?). As our results suggest, this intuition does not extend to cases in which deliberation does not necessarily lead to universal agreement, and in which deliberative resources are scarce. The informational efficiency of deliberation (which becomes the epistemic equivalent of 'truth') may be best served neither by the equal standing of speakers with respect to the deliberative procedure, nor, as the analysis below indicates, by the irrelevance of the distribution of resources available to them.

One may attempt to respond to this by arguing that providing for the possibility of hearing a dissenting opinion may have good future consequences, even if today one may find it unpersuasive (Mill, 1989). This is, of course, true to a point, but its being true relies on the possibility of delivering the reasons to those who may, in principle, be persuaded by them, and that is precisely what drives our results. Since altering the distribution of deliberative resources away from the informational optimum means failing to allow the more consequential arguments to be heard, the right question to ask, it seems to us, is not whether truth suffers from the fact that a disagreement is not aired, but which of the possible disagreements is more productively explored, given what we know about ourselves and the limitations we face.[21]

The second key implication of Proposition 3 concerns the question of what kinds of institution are most able to further that goal. In particular, the conjunction of parts (ii) and (iii) of Proposition 3 says that a (neutral) institutional designer interested in minimizing the aggregate informational loss I would be in a position to treat the problem of minimizing I as essentially unconstrained: the most informationally efficient combination of individual strategies and institutional choices is uniquely implementable in equilibrium. Moreover, as part (ii) of the proposition indicates, in determining the optimal K, we can restrict our attention to the minimization of I for each given equilibrium play, and, selecting the minimand vector K, expect to be able to implement it in that equilibrium. This is, then, the approach we now take in determining the optimality of K.[22]

Since the case of unlimited resources is often implausible, we assume that there exists a (binding) resource constraint, so that an increase in K_{A_i} must come

---

21. Of course, our results do not imply that it is beneficial for society to suppress the possibility of disagreement or make it more difficult for someone to make personal choices that are at odds with those preferred by others.

22. It is noteworthy that in all of the equilibria, agents have very different preferences over the distribution of K. The question of implementing an optimal distribution is, therefore, politically sensitive.

at the expense of a decrease in K_{A_j}, for at least some pairs of distinct types $A_i \neq A_j$. The following examples help illustrate the politically familiar nature of this choice. The first example is a consideration of an issue by a committee. The total amount of time a committee may have to spend on an issue is often fixed, and how that time is allocated often determines the nature of the discussion and its outcome. In such a context, our vector K can be interpreted as an allocation of speaking time among the members. Another, very different, application is campaign finance. In the context of the effort to reform political campaign finance in the USA, a scenario that has received considerable attention (e.g., Donnelly, Fine and Miller, 1997) goes as follows. Suppose that the supporters of the reform have been able to organize sufficient support for a measure that bans campaign contributions to the candidates, parties, or political action committees. In exchange, the government is taking upon itself the issuing of grants to the candidates to finance the sending of messages to constituents, using various media outlets, organizing campaign rallies, visits by the candidate, etc. – i.e., the sort of measures that may be thought to increase K. In the case of such a policy, it is reasonable to conclude that the resources available to candidates are finite and binding. In this case, too, the question of how to distribute the funds among the candidates acquires special importance. Our last result addresses the optimal choice of institutions when, as in these examples, institutional resources are scarce. Let $K^*_{A_i}$ be the socially (informationally) optimal value of K_{A_i}. Then:

PROPOSITION 4:

(i) $\lambda^* = (0, 1, 1, 0)$ and $K^* = (K^*_A, 0, 0, K^*_D)$, with K^*_A, $K^*_D > 0$ derived in the Appendix, are informationally optimal if both the probability of a moderate true-type is greater than the probability of each of the extreme types and the asymmetry between the radicalism of the extremists is not too great (i.e. y is neither 'too large' nor 'too small');

(ii) $\lambda^* = (0, 1, 0, 0)$ and $K^* = (K^*_A, 0, K^*_C, 0)$, with $K^*_A > 0$, $K^*_C \geq 0$ derived in the Appendix, are informationally optimal if the left-wing position is relatively more radical (i.e. if y is sufficiently small);

(iii) $\lambda^* = (0, 0, 1, 0)$ and $K^* = (0, K^*_B, 0, K^*_D)$, with $K^*_B \geq 0$, $K^*_D > 0$ derived in the Appendix, are informationally optimal if the right-wing position is relatively more radical (i.e. if y is sufficiently large).

Proof. See Appendix.

Thus, the deliberative institutions K under which the deliberative body achieves the highest equilibrium-implementable informational efficiency vary with features of the primitive features of that body, namely, with the distribution of types p and with the relative radicalism of the extreme 'right' and 'left' positions. Unless the differences between the right and moderate policy positions

and between the left and moderate positions are sufficiently similar, one of the two most inegalitarian institutions is strictly best on informational grounds – viz., one which induces only one of the two moderate types to consider reasons made by others and under which one type has no role in deliberation at all. Furthermore, when the leftist is relatively more radical than the rightist, it is informationally optimal to empower the leftists to speak and not the rightists, with the reverse being true when the rightist is relatively more radical.

This somewhat surprising result is a consequence of the neutrality of the informational criterion with respect to different individuals' true or fully-informed judgments of the best policy. Because of this neutrality, if one kind of error in individual policy choice is substantially more serious than another, the combination of institutional arrangement and equilibrium behavior that is infor-mationally optimal is the one that better eliminates the more serious error. An error may be 'serious' either because it is highly likely or because it is great in magnitude. Because the marginal utility of greater proximity to the 'correct' policy is decreasing, the expected error made by an agent who is uncertain whether she prefers the more radical extremist position or the moderate position is necessarily greater (in utility terms) than the expected error made by an agent who is uncertain whether she prefers the less radical extremist or the moderate position. The minimization of aggregate error then dictates, ceteris paribus, that the institutional environment and the deliberative behavior should be such that the agent who is compromising between the more radical extreme and moderate positions should receive the information necessary to resolve her uncertainty. Since that information includes the argument that supports exclusively the more radical extremist position, the institutional arrangement should ensure that it is heard. Since that information also includes the argument that could persuade her to adopt the moderate position, the corresponding moderate interlocutor must be induced to speak. An institutional environment that minimizes the ability of the less radical of the extremists to make her argument heard minimizes the benefits of listening accruing to the moderate interlocutor and so induces her to allocate her resources to speaking.

The optimal institutional arrangements identified in Proposition 4 may be thought to give the extreme left and right agents strong incentives to strategically exaggerate the radicalism of their policy positions. Indeed, all else equal, being able under such arrangements to appear more radical than the agent at the oppo-site policy extreme would have the effect of denying that agent the opportunity to move the moderate closest to her further away from the first agent. The possi-bility of such a strategic action is, however, severely limited. Empowering the speech of the more radical extremist is informationally optimal only because it assists a more moderate agent in resolving her uncertainty about the best policy. The fact of her uncertainty rests on the apparent plausibility, to her, of the extreme position. The more radical the extreme position, the less plausible it

becomes to the moderate agent, and so the less receptive the optimal deliberative institutions are to promulgating that radical's message.

Subject to this caveat, the results presented above on the desirability of bolstering the ability of extremists to make themselves heard are robust to expanding the set of types (possibly by introducing more arguments). For any connected mapping of this set into a unidimensional policy space, there are active argument types that correspond to specific extreme positions with certainty. Because they cannot benefit from listening themselves but, given connectedness, can and will choose to make arguments that are (in expectation) persuasive to others, the social value of empowering these extremists to speak follows.

Our final observation is a note of caution. The choice of political institutions is typically made in relation to a number of different criteria, only one of which – informational efficiency – is analyzed here. The desirability of other criteria, such as political stability, may justifiably lead us to choose institutions that are suboptimal with respect to informational efficiency. How such situations may arise may be gleaned from Proposition 4: when $\lambda^* = (0, 1, 0, 0)$ and $K^* = (K_A^*, 0, K_C^*, 0)$, or when $\lambda^* = (0, 0, 1, 0)$ and $K^* = (0, K_B^*, 0, K_D^*)$, one of the ideological extremes is, effectively, excluded from the deliberative process. To the extent that such exclusions increase the appeal to those individuals or groups of actions that are outside the legitimate channels of political influence, society may choose to sacrifice informational efficiency in the interests of discouraging them from resorting to such actions. From this standpoint, results such as those in Proposition 4 should be seen less as prescriptions than as identifications of the trade-offs we may be facing in our choices of social and political institutions, clarifying what is being sacrificed in a given institutional choice and so leading to more considered choices as a result.

Conclusion

The model in this article captures what we argue to be the key features of a common type of public deliberation – one in which the information conveyed by the speaker is accepted not on the strength of the speaker's credibility, but on the strength of the intrinsic correspondence between the propositional content of his message and the analytical structure of the current knowledge and beliefs of the listener. Since that type of deliberation is only meaningful when agents are not logically omniscient, we consider a model which relaxes logical omniscience in a way consistent with empirical evidence, but also in a way that preserves the broad contours of rational agency. Our model delivers several conclusions that both challenge the received wisdom in deliberative democratic theory and point toward a fruitful research agenda in the formal-theoretic analysis of public deliberation.

Some of the elements of this research agenda, such as the effect of deliberation on ideological group polarization, have already been pursued elsewhere (Hafer and Landa, 2005). As we showed in that article, the predictions of our model of deliberation for the ideological group dynamics in biased groups conform to the robust empirical evidence of post-deliberative group polarization (Sunstein, 2002). The model of deliberation as self-discovery holds promise for addressing several other politically important issues as well. One such issue concerns the optimal design of subgroups or subcommittees. Given the dependence – evident in our model – of the agents' relative incentives to speak or listen on the composition of the group, informational welfare is not invariant with respect to different subgroup/sub-committee divisions. For the same reasons, if speaking is sequential, we should expect variation in informational welfare depending on the order of speakers.

Another future avenue of research is the development of a theory of electoral competition that relies on what would be, arguably, a more plausible model of campaigning than that studied in the extant models. In this theory, candidates could strategically choose positions and corresponding argument supports in expectation of activating the latent arguments of the electorate and provoking endogenous political activism on their behalf from the voters who are 'more certain' of their policy positions.

Finally, the effects – in relation to various normative institutional criteria, including the one adopted in the present article – of adopting different preference aggregation rules on the practice of deliberation as self-discovery are of particular interest. A consideration of this issue will enable what promises to be an instructive comparison of the model of deliberation as self-discovery with the recent advances in the analysis of cheap-talk models of institutional effects (e.g., Austen-Smith and Feddersen, 2002; Gerardi and Yariv, 2002).

Appendix

Proof of Proposition 1: Our equilibrium is a vector $(\pi^*(\mathcal{A}_i; \mathcal{R}, \hat{\pi}(., y), p),$ $s^*(\mathcal{A}, \mathcal{R}, p, \pi^*, \lambda^*, K),\ \lambda^*(\mathcal{A}_i^0, \mathcal{A}_{-i}^0, \lambda_{-i}, \pi^*, s^*, \mathcal{R}, \hat{\pi}(., y), p, K, \theta))$ in undominated strategies, such that for all i,

$$\pi_i^*(.) \in \arg \max_{\pi_i \in [-1, y]} \sum_{\mathcal{R}_i \in \mathcal{R}} \Pr(\mathcal{R}_i | \mathcal{A}_i, \mathcal{R}, p) u(\hat{\pi}(\mathcal{R}_i), \pi_i, \pi_{-i}; .);$$

$$s_i^*(.) \in \arg \max_{s_i \in \triangle(R)} E[u(\pi^*(.) | s_i, \mathcal{A}, \mathcal{R}, p, \lambda^*)];$$

$$\lambda_i^*(.) \in \arg \max_{\lambda_i \in [0,1]} E[u_i(\hat{\pi}_i, \pi^*(\mathcal{A}_i; .); \theta, n) | s^*, p, \mathcal{R}, \hat{\pi}(., y), K, \lambda].$$

(1) π_i^* is a dominant strategy. $\pi_i^*(\{A\}) = \pi_i^*(\{A, B\})$, $\pi_i^*(\{D\}) = \pi_i^*(\{C, D\})$, and $\pi_i^*(\mathcal{R}_i) = \hat{\pi}(\mathcal{R}_i) \forall \mathcal{R}_i \in \mathcal{R}$. $\pi_i^*(\{B\})$ and $\pi_i^*(\{C\})$ maximize

$$E[u(\hat{\pi}_i, \pi)|\{B\}, .)] = -\frac{p_{AB}}{p_{AB} + p_{BC}}(\pi + 1)^2$$

$$-\frac{p_{BC}}{p_{AB} + p_{BC}}(\pi - 0)^2 + \frac{\theta}{3}\sum_{j \neq i} E[-(\hat{\pi}_i - \pi_j)^2|.]$$

$$E[u(\hat{\pi}_i, \pi)|\{C\}, .)] = -\frac{p_{BC}}{p_{CD} + p_{BC}}(\pi - 0)^2$$

$$-\frac{p_{CD}}{p_{CD} + p_{BC}}(\pi - y)^2 + \frac{\theta}{3}\sum_{j \neq i} E[-(\hat{\pi}_i - \pi_j)^2|.]$$

Solving the first-order conditions $\frac{\partial E[u(\pi)|.]}{\partial \pi} = 0$ yields the result.

(2) Recall that by assumption $\forall i \in N$, $\Pr(r \in \mathcal{L}_i|\mathcal{A}_i) \neq \emptyset \rightarrow s_{ir} = 0$. Then $\mathcal{R} \rightarrow s_{AB} = s_{BA} = s_{BC} = s_{CB} = s_{CD} = s_{DC} = 0$. From $\mathcal{R}$ and $\mathcal{A}$, $\Pr(A \in \mathcal{L}_i|\mathcal{A}_i) > 0$ if and only if $\mathcal{A}_i = \{B\}$. From (1), (2), and part 1 of this Proposition, $s'_C = (0, s_{CB}, s_{CC}, s_{CD})$ weakly dominates $s_C = (s_{CA}, s_{CB}, s_{CC}, s_{CD}) \; \forall s_{CA} > 0$; $s'_D = (0, s_{DB}, s_{DC}, s_{DD})$ weakly dominates $s_C = (s_{DA}, s_{DB}, s_{DC}, s_{DD}) \; \forall s_{DA} > 0$; and $s'_A = (s'_{AA}, s_{AB}, s_{AC}, s_{AD})$ weakly dominates $s_A = (s_{AA}, s_{AB}, s_{AC}, s_{AD}) \; \forall s'_{AA} > s_{AA}$. Therefore, $s^*_{CA} = s^*_{DA} = 0$ and $s^*_{AA} = 1 - (s_{AB} + s_{AC} + s_{AD})$. By symmetry, $s^*_{BD} = s^*_{AD} = 0$ and $s^*_{DD} = 1 - (s_{DA} + s_{DB} + s_{DC})$.

From $\mathcal{R}$ and $\mathcal{A}$, $\Pr(B \in \mathcal{L}_j|\mathcal{A}_j) > 0$ if and only if $\mathcal{A}_j = \{A\}$ or $\mathcal{A}_j = \{C\}$. From (1), (2), and part 1 of this Proposition, $s'_D = (s_{DA}, 0, s_{DC}, s_{DD})$ weakly dominates $s_D = (s_{DA}, s_{DB}, s_{DC}, s_{DD}) \; \forall s_{DB} > 0$; and $s'_B = (s_{BA}, s'_{BB}, s_{BC}, s_{BD})$ weakly dominates $s_B = (s_{BA}, s_{BB}, s_{BC}, s_{BD}) \; \forall s'_{BB} > s_{BB}$. Therefore, $s^*_{DB} = 0$ and $s^*_{BB} = 1 - (s_{BA} + s_{BC} + s_{BD})$. By symmetry, $s^*_{AC} = 0$ and $s^*_{CC} = 1 - (s_{CA} + s_{CB} + s_{CD})$.

(3) $\pi^*_i(\{A\}) = \pi^*_i(\{A, B\})$ and $\pi^*_i(\{D\}) = \pi^*_i(\{C, D\}) \Rightarrow \lambda^*_A = 0$ and $\lambda^*_D = 0$ are weakly dominant.

Suppose $\mathcal{A}_i = \{B\}$. Substituting π^*, s^*, and $\lambda^*_A = \lambda^*_D = 0$ into (2), differentiating, and gathering terms yields

$$\frac{\partial E[u(.|\{B\}, .)]}{\partial \lambda_B} = K_A \frac{p_{AB}p_{BC}^2}{(p_{AB} + p_{BC})^3} + K_C(1 - \lambda_C)\frac{p_{BC}p_{AB}^2}{(p_{AB} + p_{BC})^3}$$

$$-\frac{\theta}{3}K_B\lambda_C \frac{p_{BC}}{(p_{BC} + p_{CD})(p_{AB} + p_{BC})}$$

$$\times \left[p_{AB}\left(\left(1 + \frac{p_{CD}y}{p_{BC} + p_{CD}}\right)^2 - 1\right) + p_{BC}\left(\frac{p_{CD}y}{p_{BC} + p_{CD}}\right)^2\right]$$

The critical value of λ_C at which $\frac{\partial E[u(.|\{B\}, .)]}{\partial \lambda_B} = 0$ is

$$\tilde{\lambda}_C = \frac{3p_{AB}(p_{BC} + p_{CD})^3(p_{AB}K_C + p_{BC}K_A)}{3p_{AB}^2(p_{BC} + p_{CD})^3K_C + \theta y K_B p_{CD}(p_{BC} + p_{AB})^2(p_{CD}y(p_{AB} + p_{BC}) + p_{AB}(p_{CD}y + 2(p_{BC} + p_{CD})))}$$

Hence $\lambda^*_B = 1$ if $\lambda_C < \tilde{\lambda}_C$, $\lambda^*_B = 0$ if $\lambda_C > \tilde{\lambda}_C$, and $\lambda^*_B \in [0, 1]$ if $\lambda_C = \tilde{\lambda}_C \leq 1$.

Suppose $\mathcal{A}_i = \{C\}$. Using π^*, s^*, and $\lambda_A^* = \lambda_D^* = 0$,

$$\frac{\partial E[u(.|\{C\},.)]}{\partial \lambda_C} = K_B(1 - \lambda_B) \frac{p_{BC} p_{CD}^2 y^2}{(p_{BC} + p_{CD})^3} + K_D \frac{p_{CD} p_{BC}^2 y^2}{(p_{BC} + p_{CD})^3}$$

$$- \frac{\theta}{3} K_C \lambda_B \frac{p_{BC}}{p_{AB} + p_{BC}} \left[\frac{p_{BC}}{p_{BC} + p_{CD}} \left(\frac{p_{AB}}{p_{AB} + p_{BC}} \right)^2 \right.$$

$$\left. + \frac{p_{CD}}{p_{BC} + p_{CD}} \left(\left(y + \frac{p_{AB}}{p_{AB} + p_{BC}} \right)^2 - y^2 \right) \right]$$

The critical value of λ_B at which $\frac{\partial E[u(.|\{C\},.)]}{\partial \lambda_C} = 0$ is

$\tilde{\lambda}_B = \dfrac{3 p_{CD}(p_{AB} + p_{BC})^3 (p_{BC} K_D + p_{CD} K_B) y^2}{3 p_{CD}^2 (p_{AB} + p_{BC})^3 K_B y^2 + K_C \theta p_{AB}(p_{BC} + p_{CD})^2 (p_{AB}(p_{BC} + p_{CD}) + 2 p_{CD}(p_{AB} + p_{BC}))}$ Hence

$\lambda_C^* = 1$ if $\lambda_B < \tilde{\lambda}_B$, $\lambda_C^* = 0$ if $\lambda_B > \tilde{\lambda}_B$, and $\lambda_C^* \in [0,1]$ if $\lambda_B = \tilde{\lambda}_B \leq 1$.

$\lambda_B^* = 1$ is a best response $\forall \lambda_C$ if $\tilde{\lambda}_C > 1$. Cancelling terms and isolating $\frac{K_A}{K_B}$ yields

$$F(\theta, y, p) = \frac{\theta}{3} \frac{p_{CD}(p_{AB} + p_{BC})^2 y}{p_{AB} p_{BC}(p_{BC} + p_{CD})^3} (p_{CD}(2 p_{AB} + p_{BC}) y + 2 p_{AB}(p_{BC} + p_{CD})).$$

Similarly, $\lambda_C^* = 1$ is a best response $\forall \lambda_B$ if $\tilde{\lambda}_B > 1$. Cancelling terms and isolating $\frac{K_D}{K_C}$ yields

$$G(\theta, y, p) = \frac{\theta}{3} \frac{p_{AB}(p_{BC} + p_{CD})^2}{p_{BC} p_{CD}(p_{AB} + p_{BC})^3 y^2} (p_{AB}(p_{BC} + p_{CD}) + 2 p_{CD}(p_{AB} + p_{BC}))$$

It is straightforward to verify that the fixed points in the best-response function are those stated in the Proposition.

Proof of Proposition 2: Taking the expectation over possible consequences of deliberation, substituting the assumed communication technology, and collecting terms,

$$E[\Delta Right] = \sum_{i \in N} \sum_{r \in R} \Pr(\mathcal{L}_i = \{r\} | \mathcal{A}_i^0) \sum_{j \in N \setminus i} s_{jr}(1 - \lambda_j) K_j \lambda_i [\pi_i(\mathcal{A}_i^0 \cup \{r\}) - \pi_i(\mathcal{A}_i^0)]$$

Substituting the equilibrium choices of π and λ and simplifying, we get

$$E[\Delta Right(0,1,0,0)] = \frac{p_{AB} p_{BC}}{(p_{AB} + p_{BC})^2} (K_C - K_A)$$

$$E[\Delta Right(0,0,1,0)] = \frac{p_{BC} p_{CD}}{(p_{BC} + p_{CD})^2} y(K_D - K_B)$$

$$E[\Delta Right(0,1,1,0)] = \frac{p_{BC}p_{CD}}{(p_{BC}+p_{CD})^2}yK_D - \frac{p_{AB}p_{BC}}{(p_{AB}+p_{BC})^2}K_A$$

Solving the inequalities $E[\Delta Right(.)] > 0$ for the corresponding equilibria, we obtain the conditions under which the deliberation produces a rightward shift. The derivatives with respect to K and y are obvious and omitted.

Proof of Proposition 3: (1) Substituting π^*, s^*, λ_A^* and λ_D^*, which are common to all the equilibria in undominated strategies, into I, taking the expectation over the outcomes of deliberation, given the beliefs over the true-type profile conditioned on the known active-reasons profile, and re-arranging terms, we obtain

$$I(0,\lambda_B,\lambda_C,0) = \frac{p_{AB}p_{BC}}{(p_{AB}+p_{BC})^3}(p_{BC}(1-K_A\lambda_B)+p_{AB}(1-K_C(1-\lambda_C)\lambda_B))$$

$$+ \frac{p_{BC}p_{CD}y^2}{(p_{BC}+p_{CD})^3}(p_{CD}(1-K_B(1-\lambda_B)\lambda_C)+p_{BC}(1-K_D\lambda_C)) \quad (3)$$

$\frac{\partial^2 I(\lambda_B,\lambda_C)}{\partial \lambda_B^2}=0$ implies $\arg\min_{\lambda_B\in[0,1]}I(0,\lambda_B,\lambda_C,0)\in\{0,1\}$. Likewise $\frac{\partial^2 I(\lambda_B,\lambda_C)}{\partial \lambda_C^2}=0$ implies $\arg\min_{\lambda_C\in[0,1]}I(0,\lambda_B,\lambda_C,0)\in\{0,1\}$.

(2) Consider $\lambda^*=(0,1,1,0)$. $\frac{\partial I(\lambda^*,K,.)}{\partial K_A}<0$ and $\frac{\partial I(\lambda^*,K,.)}{\partial K_D}<0$, and $\frac{\partial I(\lambda^*,K,.)}{\partial K_B}=\frac{\partial I(\lambda^*,K,.)}{\partial K_C}=0$, so any K such that $K_A=1$ and $K_D=1$ minimizes I. $K=(1,0,0,1)$ corresponds to the upper right quadrant of Figure 1, and hence, from Proposition 1, implements $(0,1,1,0)$ as the unique equilibrium. By a similar argument, given $\lambda^*=(0,1,0,0)$, $K=(1,0,1,0)$ minimizes I and uniquely implements λ^*, and given $\lambda^*=(0,0,1,0)$, $K=(0,1,0,1)$ minimizes I and uniquely implements λ^*.

(3) Follows from (1) and (2).

Proof of Proposition 4: Let $\kappa^*(\lambda,k,p,y)\in\arg\min_{K\in\{K|\sum_{i\in N}K_i\leq k\}}I(\lambda,K,p,y)$. Given the Proposition 3, we need only to compare $I(\lambda,\kappa^*(\lambda,.),.)$ for $\lambda\in\{(0,1,1,0),(0,1,0,0),(0,0,1,0)\}$ to identify

$$K^*(k,p,y)\in\arg\min_{K\in\{K|\sum_{i\in N}K_i\leq k\}}I(\lambda^*(K,.),K,p,y).$$

Henceforth, let $\lambda^{*1}=(0,1,1,0)$, $\lambda^{*2}=(0,1,0,0)\%$, and $\lambda^{*3}=(0,0,1,0)$.

We first describe $\kappa^*(\lambda,k,p,y|\lambda)$. Substitute λ^{*j} into (3), and let $Q(\lambda^{*j})=\{r|r\in R$ and $\frac{\partial I(\lambda)}{\partial K_r}<0\}$ and index its elements $i=1,2$ such that $\frac{\partial I(\lambda)}{\partial K_{q_1}}\leq\frac{\partial I(\lambda)}{\partial K_{q_2}}$. Then $\kappa_{q_1}^*(\lambda^{*j},k,p,y)=\min\{k,1\}$, $\kappa_{q_2}^*(\lambda^{*j},k,p,y)=\max\{0,\min\{k-1,1\}\}$, and $\forall r\notin Q(\lambda^{*j})$, $\kappa_r^*(\lambda^{*j},k,p,y)=0$. Letting $X=\frac{(p_{BC}+p_{CD})^3 p_{AB}}{(p_{AB}+p_{BC})^3 p_{CD}}$,

$$\kappa^*(\lambda^{*1}, k, p, y) = \begin{cases} (1,0,0,1) & \text{if } k \geq 2 \\ (1,0,0,k-1) & \text{if } k \in [1,2) \text{ and } X > y^2 \\ (k,0,0,0) & \text{if } k < 1 \text{ and } X > y^2 \\ (k-1,0,0,1) & \text{if } k \in [1,2) \text{ and } X < y^2 \\ (0,0,0,k) & \text{if } k < 1 \text{ and } X < y^2. \end{cases}$$

$$\kappa^*(\lambda^{*2}, k, p, y) = \begin{cases} (1,0,1,0) & \text{if } k \geq 2 \\ (1,0,k-1,0) & \text{if } k \in [1,2) \text{ and } p_{BC} > p_{AB} \\ (k,0,0,0) & \text{if } k < 1 \text{ and } p_{BC} > p_{AB} \\ (k-1,0,1,0) & \text{if } k \in [1,2) \text{ and } p_{BC} < p_{AB} \\ (0,0,k,0) & \text{if } k < 1 \text{ and } p_{BC} < p_{AB} \end{cases}$$

$$\kappa^*(\lambda^{*3}, k, p, y) = \begin{cases} (0,1,0,1) & \text{if } k \geq 2 \\ (0,1,0,k-1) & \text{if } k \in [1,2) \text{ and } p_{CD} > p_{BC} \\ (0,k,0,0) & \text{if } k < 1 \text{ and } p_{CD} > p_{BC} \\ (0,k-1,0,1) & \text{if } k \in [1,2) \text{ and } p_{CD} < p_{BC} \\ (0,0,0,k) & \text{if } k < 1 \text{ and } p_{CD} < p_{BC} \end{cases}$$

By comparing $I(\lambda^{*1}, \kappa^*(\lambda^{*1},.),.)$, $I(\lambda^{*2}, \kappa^*(\lambda^{*2},.),.)$, and $I(\lambda^{*3}, \kappa^*(\lambda^{*3},.),.)$, we identify for each $j = 1, 2, 3$ the conditions under which $I(\lambda^{*j}, \kappa^*(\lambda^{*j},.),.)$ is least, and hence identify $K^*(k, p, y)$.

(1) $k \geq 2$. $\lambda^{*1} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, 2,.),.)$, and hence $K^*(.) = (1,0,0,1)$ iff $y^2 \in (X\frac{p_{AB}}{p_{BC}}, X\frac{p_{BC}}{p_{CD}})$. If $y^2 < X\frac{p_{AB}}{p_{BC}}$, then $K^*(.) = (1,0,1,0)$; and if $y^2 > X\frac{p_{BC}}{p_{CD}}$, then $K^*(.) = (0,1,0,1)$.

(2) $k \in (1,2)$. For each of the three equilibria, $\kappa^*(\lambda, .)$ has two possible values conditional on (p, y), generating eight subcases:

 (i) $X > y^2$ and $p_{CD} > p_{BC} > p_{AB}$. Then $\lambda^{*1} = \arg\min I(\lambda^*, \kappa^*(\lambda^*,.),.)$, and hence $K^*(.) = (1,0,0,k-1)$, iff $y^2 \in (X\frac{p_{AB}}{p_{BC}}, X\frac{p_{BC}}{p_{CD}})$. If $y^2 < X\frac{p_{AB}}{p_{BC}}$, then $K^*(.) = (1,0,k-1,0)$; and if $y^2 > X\frac{p_{BC}}{p_{CD}}$, then $K^*(.) = (0,1,0,k-1)$.

 (ii) $X < y^2$ and $p_{CD} < p_{BC} < p_{AB}$. Then $\lambda^{*1} = \arg\min I(\lambda^*, \kappa^*(\lambda^*,.),.)$, and hence $K^*(.) = (k-1,0,0,1)$, iff $y^2 \in (X\frac{p_{AB}}{p_{BC}}, X\frac{p_{BC}}{p_{CD}})$. If $y^2 < X\frac{p_{AB}}{p_{BC}}$, then $K^*(.) = (k-1,0,1,0)$; and if $y^2 > X\frac{p_{BC}}{p_{CD}}$, then $K^*(.) = (0,k-1,0,1)$.

 (iii) $X > y^2$ and $p_{BC} > p_{AB}$, $p_{CD} < p_{BC}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k,.),.) \leq I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k,.),.)$. Then if $y^2 < X\frac{p_{AB}}{p_{BC}}$, then $K^*(.) = (1,0,k-1,0)$; else $K^*(.) = (1,0,0,k-1)$.

 (iv) $X < y^2$ and $p_{BC} > p_{AB}$, $p_{CD} < p_{BC}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k,.),.) \leq I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k,.),.)$. Then if $y^2 < X\frac{p_{BC}}{p_{CD}}$, then $K^*(.) = (k-1,0,0,1)$; else $K^*(.) = (0,k-1,0,1)$.

(v) $X > y^2$ and $p_{CD} < p_{BC} < p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) < I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$ and $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$. Then $\lambda^{*2} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (k-1, 0, 1, 0)$.

(vi) $X < y^2$ and $p_{CD} > p_{BC} > p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) < I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$ and $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$. Then $\lambda^{*3} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, 1, 0, k-1)$.

(vii) $X > y^2$ and $p_{BC} < p_{AB}$, $p_{CD} > p_{BC}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$. If also $p_{AB} > p_{CD}$, then $\lambda^{*2} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (k-1, 0, 1, 0)$. If $p_{AB} < p_{CD}$, then $K^*(.) = (k-1, 0, 1, 0)$ if $y^2 < X \frac{(p_{AB} + p_{BC}(k-1))}{(p_{CD} + p_{BC}(k-1))}$, else $\lambda^{*3} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$ and hence $K^*(.) = (0, 1, 0, k-1)$.

(viii) $X < y^2$ and $p_{CD} > p_{BC}$, $p_{BC} < p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$. If also $p_{AB} < p_{CD}$, then $\lambda^{*3} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, 1, 0, k-1)$. If $p_{AB} \geq p_{CD}$, then $K^*(.) = (0, 1, 0, k-1)$ if $y^2 > X \frac{(p_{AB} + p_{BC}(k-1))}{(p_{CD} + p_{BC}(k-1))}$, else $\lambda^{*2} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$ and hence $K^*(.) = (k-1, 0, 1, 0)$.

(3) $k \in (0, 1]$. Again, there are eight subcases:

(i) $X > y^2$ and $p_{CD} > p_{BC} > p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) = I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$. If $y^2 < X \frac{p_{BC}}{p_{CD}}$, then $I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .) < I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$, and hence $K^*(.) = (k, 0, 0, 0)$; else, $K^*(.) = (0, k, 0, 0)$.

(ii) $X < y^2$ and $p_{CD} < p_{BC} < p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) = I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$. If $y^2 > X \frac{p_{AB}}{p_{BC}}$, then $I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .) < I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$, and hence $K^*(.) = (0, 0, 0, k)$; else, $K^*(.) = (0, 0, k, 0)$.

(iii) $X > y^2$ and $p_{BC} > p_{AB}$, $p_{CD} < p_{BC}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) = I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .) < I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$. Hence $K^*(.) = (k, 0, 0, 0)$.

(iv) $X < y^2$ and $p_{BC} > p_{AB}$, $p_{CD} < p_{BC}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) = I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .) < I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$. Hence $K^*(.) = (0, 0, 0, k)$.

(v) $X > y^2$ and $p_{CD} < p_{BC} < p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$ and $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) < I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$. Then $\lambda^{*2} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, 0, k, 0)$.

(vi) $X < y^2$ and $p_{CD} > p_{BC} > p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$ and $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) < I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$. Then $\lambda^{*3} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, k, 0, 0)$.

(vii) $X > y^2$ and $p_{BC} < p_{AB}$, $p_{CD} > p_{BC}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*2}, \kappa^*(\lambda^{*2}, k, .), .)$. If also $p_{AB} > p_{CD}$, then $\lambda^{*2} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, 0, k, 0)$. If $p_{AB} < p_{CD}$, then $\lambda^{*3} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, k, 0, 0)$ if $y^2 > X \frac{p_{AB}}{p_{CD}}$, else $K^*(.) = (0, 0, k, 0)$.

(viii) $X < y^2$ and $p_{CD} > p_{BC}$, $p_{BC} < p_{AB}$. Then $I(\lambda^{*1}, \kappa^*(\lambda^{*1}, k, .), .) \geq I(\lambda^{*3}, \kappa^*(\lambda^{*3}, k, .), .)$. If also $p_{AB} < p_{CD}$, then $\lambda^{*3} = \arg\min I(\lambda^*, \kappa^*(\lambda^*, .), .)$, and hence $K^*(.) = (0, k, 0, 0)$. If $p_{AB} > p_{CD}$, then $\lambda^{*2} = $ arg min

$I(\lambda^*, \kappa^*(\lambda^*,.),.)$,   and   hence $K^*(.) = (0,0,k,0)$ if $y^2 < X\frac{p_{AB}}{p_{CD}}$,   else $K^*(.) = (0,k,0,0)$.

REFERENCES

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite and David Schmeidler (2001) 'Rhetoric and Analogies', Mimeo, Tel-Aviv University.

Austen-Smith, David (1990) 'Information Transmission in Debate', *American Journal of Political Science* 34(1): 124–52.

Austen-Smith, David and Timothy Feddersen (2002) 'Deliberation and Voting Rules', Mimeo, Northwestern University.

Baron, David P. (2003) 'Private Politics', *Journal of Economics & Management Strategy* 12(1): 31–66.

Baron, Jonathan (1994) *Thinking and Deciding*. Cambridge: Cambridge University Press.

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

Calvert, Randall and James Johnson (1998) 'Rational Actors, Political Reasons and Democratic Deliberation', Mimeo, University of Rochester.

Christiano, Thomas (1997) 'The Significance of Public Deliberation', in J. Bohman and W. Rehg (eds) *Deliberative Democracy: Essays on Reason and Politics*, pp. 243–78. Cambridge: MIT Press.

Cohen, Joshua (1997) 'Deliberation and Democratic Legitimacy', in J. Bohman and W. Rehg (eds) *Deliberative Democracy: Essays on Reason and Politics*, pp. 67–92. Cambridge: MIT Press.

Crawford, Vincent and Joel Sobel (1982) 'Strategic Information Transmission', *Econometrica* 50, 1431–51.

Dawes, Robyn M. (1998) 'Behavioral Decision Making and Judgment', in D. T. Gilbert et al. (eds) *The Handbook of Social Psychology*, 4th ed., pp. 497–548. Boston, MA: McGraw Hill.

Dickson, Eric, Catherine Hafer and Dimitri Landa (2005) 'Cognition and Strategy: A Deliberation Experiment', Mimeo, New York University.

Donnelly, David, Janice Fine and Ellen Miller (1997) 'Going Public', *Boston Review* 22(2)

Elster, Jon (ed) (1998) *Deliberative Democracy*. Cambridge: Cambridge University Press.

Fearon, James D. (1998) 'Deliberation as Discussion', in J.Elster (ed.) *Deliberative Democracy*, pp. 44–68. Cambridge: Cambridge University Press.

Føllesdal, Dagfinn (1982[1994]) 'The Status of Rationality Assumptions in Interpretation and Explanation of Action', *Dialectica* 36: 301–16.

Geanakoplos, John (1989) *Games Theory Without Partitions, and Applications to Speculation and Consensus*, Cowles Foundation Discussion Paper No. 914.

Gerardi, Dino and Leeat Yariv (2002) 'Putting Your Mouth Where Your Mouth Is: An Analysis of Collective Choice with Communication', Mimeo, Yale University.

Goldberg, Jonah (2003) 'Springfield vs. Shelbyville: Gay Marriage, Incest, and *The Simpsons*', National Review Online, July 1 2003. Available at http://www.nationalreview.com/goldberg/goldberg070103.asp

Gutmann, Amy and Dennis F. Thompson (1996) *Democracy and Disagreement.* Cambridge, MA: Harvard University Press.

Habermas, Jürgen (1989) *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society.* Cambridge, MA: MIT Press.

Habermas, Jürgen (1996) 'Popular Sovereignty as Procedure', in J. Habermas and W. Rehg, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, pp. 491–515. Cambridge, MA: MIT Press.

Hafer, Catherine and Dimitri Landa (2005) 'Deliberation, Ideological Bias, and Group Choice', Mimeo, New York University.

Hurley, Susan (2000) 'Cognitivism in Political Philosophy', in R. Crisp and B. Hooker (eds) *Well-Being and Morality*, pp. 177–209. Oxford: Clarendon Press.

Johnson, James (1998) 'Arguing for Deliberation: Some Skeptical Considerations', in J. Elster (ed) *Deliberative Democracy*, pp. 161–84. Cambridge: Cambridge University Press.

Kennedy, Randall (1999) 'Racial Profiling Usually Isn't Racist. It Can Help Stop Crime. And It Should Be Abolished', *New Republic*, September 13, 1999.

Landa, Dimitri (2006) 'Rational Choices as Social Norms', *Journal of Theoretical Politics* 18(4): 434–53.

Lord, C. G., L. Ross and M. R. Lepper (1979) 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence', *Journal of Personality and Social Psychology* 37: 2098–109.

Lupia, Arthur (2002) 'Deliberation Disconnected: What It Takes to Improve Civic Competence', *Law and Contemporary Problems* 65(3): 133–50.

Macedo, Stephen (ed.) (2000) *Deliberative Politics: Essays on Democracy and Disagreement*. Oxford: Oxford University Press.

Mackie, Gerry (1998) 'All Men Are Liars: Is Democracy Meaningless?' in J. Elster (ed.) *Deliberative Democracy*, pp. 69–96. Cambridge: Cambridge University Press.

Manin, Bernard (1987) 'On Legitimacy and Political Deliberation', *Political Theory* 15(3): 338–68.

Meirowitz, Adam (2007) 'In Defense of Exclusionary Deliberation: Communication and Voting with Private Beliefs and Values', *Journal of Theoretical Politics* 19(3): 301–27.

Mill, John Stuart (S. Collini, ed.) (1989) *On Liberty*. Cambridge: Cambridge University Press.

Rabin, Matthew (1998) 'Psychology and Economics', *Journal of Economic Literature* 36 (March): 11–46.

Rabin, Matthew and Joel L. Schrag (1999) 'First Impressions Matter: A Model of Confirmatory Bias', *The Quarterly Journal of Economics* (February): 37–82.

Rawls, John (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Sunstein, Cass (2002) 'The Law of Group Polarization', *Journal of Political Philosophy* 10(2): 175–95.

Tetlock, Philip E. (1992) 'The Impact of Accountability on Judgment and Choice: Toward a Social Contingency model', Advances in Experimental Social Psychology 25: 331–76.

Wason, P. C. (1968) 'Reasoning about a Rule', *Quarterly Journal of Experimental Psychology* 20: 273–81.

Wason, P. C. (1977) 'Self-Contradictions', in P. N. Johnson-Laird and P. C. Wason (eds) *Thinking: Readings in Cognitive Science*, pp. 114–28. Cambridge: Cambridge University Press.

Wood, Genevieve (2003) National Public Radio Morning Edition, 4 September 2003.

Zaller, John (1992) *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

CATHERINE HAFER is an Assistant Professor of Politics at New York University. She has published work on regulatory politics, property rights, and deliberation in the *American Political Science Review, Review of Economic Studies, Journal of Politics*, and other journals. Her current work focuses on three main projects: deliberation in politics, the relationship between regulation and lobbying, and political conflict and mediation.

DIMITRI LANDA is an Assistant Professor of Politics at New York University. His work on various aspects of social, political, and legal theory has appeared in

American Political Science Review, Journal of Politics, Journal of Theoretical Politics, Law and Philosophy, and other journals and edited volumes. His current projects include work on deliberative institutions, on the informational effects of strategic challenger entry in elections, and on political accountability and electoral representation. [email: dimitri.landa@nyu.edu].