Deliberation and Social Polarization*

Catherine Hafer† Dimitri Landa‡

01/25/2006, first draft: 05/16/2005

Abstract

We develop a theory of social polarization induced by "deliberation as self-discovery." In such deliberation, intrinsically persuasive arguments activate the "latent" reasons of the corresponding listeners, whose beliefs about the best alternative change only in response to arguments they find persuasive. In equilibrium, agents sort into ideologically biased groups with speakers whose ideological bias reinforces their own. These choices, in turn, give rise to group polarization - a widely cited phenomenon whereby deliberation in biased groups leads individuals to adopt post-deliberative positions more extreme than their prior bias.

---

*This paper was initially circulated under the title "Deliberation as Self-Discovery and Group Polarization." We have benefitted from comments from Ethan Bueno de Mesquita, Sanford Gordon, Rebecca Morton, and the participants at the Kellogg School of Business Conference on Deliberation and Collective Choice (May 2005).

†Assist. Prof. of Politics, NYU, e-mail: catherine.hafer@nyu.edu

‡Assist. Prof. of Politics, NYU, e-mail: dimitri.landa@nyu.edu

I. Introduction

The formation of deliberative groups is one of the central features of democratic politics and, increasingly, a focus of positive analyses of collective decision-making (Murphy and Shleifer [2004]; Glaeser, Ponzetto, and Shapiro [2005]). The politically relevant interactions within such groups occur in a wide range of settings, including collegial courts and legislative committees, informal political or religious associations, and anonymous internet-based chatrooms. As an empirical matter, the causes and consequences of such interactions are often context-specific, but there are also general regularities that appear to hold for considerable classes of contexts. Some of the best-documented such regularities concern two key questions regarding the macro-level political consequences of group deliberation: (1) given free association, what kinds of deliberative groups of agents should be be expected to form? and (2) what aggregate changes in individuals' policy choices may be expected as a result of deliberation in groups of various ideological compositions?

A key finding with respect to the first question is that individuals tend to join groups of like-biased individuals: those who are ideologically right-leaning tend to join rightist deliberative groups, while those who are left-leaning the leftist ones (Shapiro [1999]; Sunstein [2002]). The answer to the second question is a well-documented phenomenon of post-deliberative group polarization - the tendency for members of ideologically biased groups to become still more extreme in the direction of the prior bias as a consequence of group deliberation (Moscovici and Zavalloni [1969]; Burnstein and Vinokur [1977]; Lindzey and Aronson [1985]; Mendelberg [2002]; Sunstein [2001, 2002]). The conjunction of these two observations suggests that factors responsible for making free association in democratic societies more effective ensure that the deliberative process deepens social polarization.

We present a informational theory of deliberation-induced social polarization that provides an explanation for this outcome and for the two observations on which it relies. A key element of our explanation is a theory of instrinsically persuasive argumentation in deliberation. Much of what goes on in deliberative settings is people explaining to other people the logical consequences of their beliefs (Calvert and Johnson [1998]; Hafer and Landa [forthcoming]). For instance, a fundamentalist Christian might point out to another self-professed believer that, if they are faithful to the Bible, then they must believe that life begins at conception, and thus must be pro-life. A person who is not persuaded by such an argument doesn't argue with the logic, they argue with the foundational assumptions (e.g., that every word of the Bible is true). Thus, much of what is going on in deliberation is not the transmission of new or private information about facts or states of the world, but the teasing out of arguments (or the reminding of precedents - as in Aragones et al. [2001]). In such instances, the speaker is convincing because, for the relevant listeners, her argument is intrinsically persuasive, rather than because she is perceived a credible source of information.

Deliberation that relies on that kind of argumentation (which we refer to as *deliberation as self-discovery*) poses a substantial theoretical problem for the standard account of rational choice. That account assumes, in effect, that people are logically omnicient, that is, that they "automatically" understand all of the logical implications of their knowledge, beliefs, and preferences. But if this is so, then there is no room for deliberation that involves one person teasing out for another the logical implications of some belief. Yet, as described above, this kind of deliberation is the modus operandi of many politically relevant deliberative

exchanges, and so developing an internally consistent account of it requires a somewhat different cognitive model of individual agency (Hafer and Landa [forthcoming]).[1]

Our model of deliberation as self-discovery combines intrinsically persuasive argumentation with an empirically plausible model of agency that could give rise to this kind of deliberation. Different agents find different reasons or arguments compelling, and these reasons can affect their beliefs about and choices of policy. For each agent, some of these compelling reasons are *active*: she knows that these reasons are both valid and relevant, and they determine her perceived optimal choice. Other of these compelling reasons are *latent*: she is initially unsure of either their validity or relevance, but will be convinced of both when she hears arguments that correspond to them. Hearing such an argument "activates" the corresponding latent reason, turning it into an active one.

For the agents in our model learning requires (direct) exposure to their latent reasons. In essence, these agents confound validity and relevance: they fail to make the possible policy-relevant inferences from messages that do not correspond to either their active or latent reasons (i.e., from arguments they consider to be invalid). In the absence of messages that directly activate their latent reasons (i.e., arguments they do consider to be valid), they make their evaluations of policy alternatives on the basis of their prior beliefs alone. The relaxation of logical omniscience entailed in this kind of updating is axiomatically well-defined and has strong experimental support (see below). The agency it describes is, moreover, naturally consistent with systematic observations of the occurrence of deliberation as self-discovery: if agents update their beliefs only in response to direct arguments, i.e., arguments that draw logical inferences from propositions they already hold true, then their failure to encounter some such arguments explains their interest in this kind of deliberation.

The heart of our model is the underlying connection between individual learning and the composition of deliberative groups that may be expected to form in equilibrium. Individuals who update their beliefs as described above can be shown to have a preference for hearing the most ideologically extreme speech that is biased in the same direction as their own prior policy position. Given free association, this has the consequence of encouraging the creation of deliberative groups that consist of agents who are, in expectation, biased in the same ideological direction, and that produce the most (ideologically) extreme speech that is consistent with their bias. Because those speakers are ideologically extreme representatives of the same bias, we expect to observe a greater prevalence of arguments that move individuals in the direction of their *ex ante* bias, given their cognitive agency. In equilibrium, then, the listeners hear the speech that reinforces their prior bias and moves them further in the direction of that bias.

We show that this result holds uniformly when agents' knowledge of each other is fully induced by their summary ideology and that it may, to a limited degree, be attenuated when they have better knowledge of the scope of reasons that may be latently held by others and can thus target more finely their speech on particular issue/argument dimensions in expectation of listener self-selection. We show, further, that sufficiently biased agents may develop

[1] Recognizing the pervasiveness of phenomena like deliberative self-discovery, Austen-Smith and Feddersen note that "models permitting failures of logical, as well as informational, omniscience are going to prove important" [2002, p. 36].

a preference against hearing (potentially persuasive) ideologically contrary arguments, even when hearing them carries no opportunity cost, and that we should expect the appeal of speaking, as opposed to listening, to be greatest for the most ideologically biased agents.

Both the nature and the scope of our explanation for social polarization stand in contrast to the existing explanations. The latter, which focus on what happens in biased groups with fixed membership, turn on two mechanisms: (1) social and peer pressures, which rely on individual susceptibility to group influence (Sunstein [2002]); (2) the ideological effects of exposure to a one-sided argumentation (Burnstein and Vinokur [1977]). Both of these mechanisms, however, fall short of what a convincing account of group polarization would require.

Although the social and peer pressure explanation accounts, no doubt, for some of the polarization findings, it does not explain why the ideological effects of group deliberation persist after the individuals leave the setting of the group discussion. The second mechanism, responsiveness to one-sided exposure, taken by itself begs the question about microfoundations: if agents know that they belong to biased groups, should they not discount the biased arguments generated by them? Moreover, if the explanation for group polarization is "informational,"[2] then the informational advantages of belonging to one group or another ought to figure in the individual choice of group membership - a causal connection that has been outside the scope of the existing explanations. On the other hand, fully rational agents have been shown to derive informational advantages from hearing a speaker whose bias is like their own rather than an unbiased speaker (Crawford and Sobel [1982]; Calvert [1985]), but the corresponding models deal with cheap-talk rather than intrinsically persuasive arguments and do not predict that the expected movement in the listener's position will be in the direction of that listener's initial bias (and so do not produce group polarization).

The remainder of the paper proceeds as follows. In Section II, we introduce the informational framework of deliberation as self-discovery. Section III presents our central results on deliberation in groups, starting with the baseline null result for Bayesian agents and then developing the model of individual choice of groups and speech that leads to group bias and group polarization. Section IV concludes with a brief discussion. The appendix in Section V gathers proofs of all the formal results in the paper.

II. 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 analysis of deliberation in groups developed in the following sections. The existing formal literature on deliberation focuses, by and large, on the analysis of cheap-talk models of information transmission (Crawford and Sobel [1982]; Ger-

[2] An explanation akin to informational cascades may also be plausible in some instances, but its relevance is limited by the nature of the arguments aired in the groups. Because that mechanism relies on agents having and believing others to have private information about the merits of the relevant alternatives, it simply does not fit the case of deliberation with intrinsically persuasive argumentation.

ardi and Yariv [2002]; Austen-Smith and Feddersen [2004]; Meirowitz [2004]). Lipman and Seppi [1995] and Lanzi and Mathis [2004] analyze a sender-receiver model in which senders can supply partial proofs for their signals and in which the veridicality (truth content) of a message is the same for all types of receivers. Common veridicality is also assumed in Glazer and Rubinstein [2005], who analyze a model of persuasion in which the messages contain "hard" or fully provable information, but this information is partial and the informativeness of messages is a function of speaker credibility. By contrast, in the model we analyze, messages are fully provable and complete, yet their veridicality differs across agents. The most natural examples of this setting are precisely the messages in which arguments are of the "self-discovery" nature described in the Introduction.

In order to operationalize i's uncertainty over her ideal policy in a manner consistent with the notion of deliberation discussed above, suppose that i's beliefs about her ideal policy choice are a function of the reasons she finds convincing. For each individual (or agent) i, the list of such convincing reasons is $r_i \in \{0,1\}^n$, where the jth element of that list , r_i^j, $j = 1, ..., n$ is an independent draw of a Bernoulli random variable with $\Pr(r_i^j = 1) = \theta_i \in [0,1]$. The value of r_i^j can be interpreted as signaling how convinced a given person is by opposing arguments with respect to dimension j, with $r_i^j = 1$ interpreted as i being convinced by a "right" argument and $r_i^j = 0$ as i being convinced by a "left" argument with respect to that dimension. (Because the value of θ_i is agent-specific and each agent's list of reasons r_i is independently drawn, different members of a society will not find the same arguments compelling.) The vector r_i can be thought of as i's type and θ_i as i's summary ideological characteristic, determining how likely i is to be convinced by "right" (or, complementarily, "left") arguments. We suppose that each i is uncertain about θ_i, but that the distribution of θ_i, described by the probability density function $p(\theta)$, is common knowledge.

Nature randomly selects and reveals some dimensions of r_i. Let $\mathcal{A}_i' \subseteq \{1, 2, ..., n\}$ contain the dimensions of the reasons that Nature reveals so that $j \in \mathcal{A}_i'$ if and only if i knows r_i^j. Call r_i^j *active* if $j \in \mathcal{A}_i'$, and *latent* if $j \notin \mathcal{A}_i'$ and let $\mathcal{L}_i'$ be the set of all such latent dimensions, so that $(\mathcal{A}_i', \mathcal{L}_i')$ is a partition of the set of dimensions $\{1, 2, ..., n\}$. Let l_i' be the number of those revealed dimensions on which $r_i^j = 1$ and let m_i' be the number of those revealed dimensions on which $r_i^j = 0$, so that $|\mathcal{A}_i'| = l_i' + m_i'$.

When referring to arguments r^j, 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 the 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 both the argument itself and its label as r^j, with the understanding that, in a general case, agents may, prior to deliberation, know only a label, and not whether they are convinced by its corresponding argument (unless the reason expressed by that argument is already active). Whether a heard argument s^j is one's latent argument or not depends on whether one would accept the set of premises it employs. Thus, when referring to r_i^j as i's latent reason, we mean that there is a set of premises that i accepts

that would, on examination, imply to her the validity of the thesis r_i^j.

From the common knowledge distribution of θ_i, players derive common conditional probabilities $\Pr(r_i|\mathcal{A}_i')$. No new information about the primitive probabilities $\Pr(r_i)$ is available from playing the game. Although r_i cannot change, $\mathcal{A}_i'$ can. Let s be the vector of revealed reasons. If agent i receives an argument s^j that matches her type r_i^j, it becomes (or stays) active; if s^j does not match r_i^j, then she does not recognize the argument and no change in $\mathcal{A}_i'$ occurs. This cognitive capacity is summarized formally as

$$\mathcal{A}_i = \{j : j \in \mathcal{A}_i' \text{ or } r_i^j = s^j\}. \tag{1}$$

This model is agnostic with respect to the nature of the inference entailed in the recognition of one's own latent argument. One type of inference consistent with (1) could be represented as 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 an inductive structure of belief updating may be modeled by the recognition of latent reasons as well.

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 that are not actively available to them prior to deliberation for developing the corresponding policy position. 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.[3] Upon hearing an argument that is not their "latent" argument, 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 arguments they must or are likely to agree with, and so will fail to update their beliefs about their induced ideal policies accordingly.

One can interpret this behavioral assumption as saying that agents 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 arguments. The "sound and valid propositional support" is a valid argument that proceeds from the premises that are shared by the listener. The listener's "latent" reason may be thought of as precisely that argument.

Dickson, Hafer, and Landa [2005] present direct experimental support for our cognitive assumption on agency in the strategic environment. There is an extensive literature supplying indirect evidence as well. One of the most robust experimental findings in cognitive and social psychology is the existence of a cognitive bias in favor of one's own currently held convictions - whereby, in order to trigger a change or a reversal in one's prior position, the argument must be direct and leave little to ambiguity (Tetlock [1992]; Zaller [1992]; Dawes

[3]We show that their deviation from the Bayesian inference is, in fact, reducible to their violation of negative introspection.

[1998]; Rabin [1998]; Baron [1994]). 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]). Perhaps most strikingly, studies of hypothesis testing (Wason [1968, 1977]; Baron [1994, Ch. 13]) find systematic evidence of reluctance to see even the possibility of making valid inferences from disconfirmations of the consequent.

III. Deliberation in Groups

The Model

We begin by assuming that the population is divided into potential speakers and potential listeners. We then consider an extension in which each agent is able to choose whether to speak or listen, and, in that context, demonstrate the substantive robustness of the results obtained in the first model. Let $\mathcal{S}$ be the set of senders/speakers and $\mathcal{R}$ the set of receivers/listeners. For each agent i, we assume that $\theta_i \in [0,1]$, which we described as i's summary ideological characteristic, is also i's ideal point in the underlying policy space. Each receiver chooses a policy $\pi_i \in [0,1]$.

A player i's (true) type θ_i is not observable, even to herself. To keep things simple, we assume that $\theta_i \sim U[0,1]$, i.e., that player i's θ_i is randomly drawn from the uniform distribution on interval $[0,1]$. Recall that l_i' is the number of dimensions randomly revealed by Nature at the beginning of play on which $r_i^j = 1$ and m_i' is the number of those revealed dimensions on which $r_i^j = 0$. As a matter of interpretation, we suppose that potential speakers' reasons are active, i.e., that $\mathcal{A}_i = \{1, 2, ..., n\}$ and thus $l_i' + m_i' = n \; \forall i \in \mathcal{S}$. The speakers' ability to make potentially persuasive arguments on every issue dimension follows naturally from such a supposition, however, in the technical sense it plays no role in establishing the results that follow and thus is, strictly speaking, unnecessary. (For the sake of clarity, we also dispense with the possibility that speakers sometimes make arguments of poor quality that are not persuasive to anyone. However, it it straightforward to show that the substantive implications of the resuts below are robust to assuming that with some fixed probability each speaker's argument on any dimension may be "bad" in this sense.)

Note that knowledge of r_i does not imply certain knowledge of θ_i. Accordingly, whether i is a speaker or a receiver, pre-deliberative beliefs about her own type θ_i can be characterized by the probability density function

$$p(\theta | l', m') = \frac{\theta^{l'}(1-\theta)^{m'}}{\int_0^1 \hat{\theta}^{l'}(1-\hat{\theta})^{m'} d\hat{\theta}}. \tag{2}$$

We assume that speakers observe the initial biases l', m' of each of the receivers.

The main sequence of play we analyze is as follows. First, each speaker chooses whether to speak or be silent. If speaker i chooses to speak, then she also chooses and makes public her speech parameter $\sigma_i \in [0,1]$, such that each of her speech arguments s_i^j, $j = 1, ..., n$ is an independent draw of a Bernoulli random variable with $\Pr(s_i^j = 1) = \sigma_i$. Next each receiver chooses/joins a speaker, knowing that her speech is characterized by σ_i. Then, each speaker

i who chooses to speak addresses the receivers who join her and makes an n-dimensional vector of arguments s_i, which is generated from σ_i in the manner described above.[4] (The stochastic speech realization means that this game form may be interpreted as substantively equivalent to assuming that speakers' and listeners' knowledge of each other is fully induced by their expected summary ideological parameters, $E[\theta_i]$.) The receivers hear the arguments of their respective speakers, and make their policy choices in accordance with their updated beliefs.[5]

Rather than explicitly modeling preference aggregation, we assume that agents' payoffs depend directly on the policy choices of the 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.[6]

Each receiver i chooses a policy $\pi_i \in [0,1]$, and her utility from that choice is $-(\pi_i - \theta_i)^2$. The quadratic functional form is the standard means of capturing risk aversion in the spatial models of policy choice. The substantive implications of our results extend to any concave function. Because the receiver cannot affect the policy choices of others, and hence her utility from their choices, her behavior is assumed to maximize her expected utility from her own choice, $-\int_0^1 p(\theta|l_i', m_i')(\pi_i - \theta)^2 d\theta$. Because speakers do not obtain new information in the course of deliberation, and hence their beliefs and the consequent policy choices remain constant with respect to their deliberative choices, each speaker i's behavior is assumed to maximize her expected utility from the choices of the receivers, $-\int_0^1 p(\theta|l_i', m_i')(\sum_{j\in\mathcal{R}}(\pi_j - \theta)^2) d\theta$.

We assume that all agents choose their actions to maximize their expected utility, given their beliefs at the time of action and the strategies of other players. Additionally, we assume that their beliefs change in accordance with the cognitive model described in the previous

---

[4] The speaker who announces that her speech is σ is, thus, committed to making that speech when listeners join the group. In a repeated setting not modeled here, groups or speakers can be seen to want to develop and maintain reputations for their degree of conservativism/liberalism in order to be able to attract audience.

[5] Before analyzing this specification of the game form, we also consider a variation of it in which the receivers have no opportunity to choose a specific speaker, and all speakers who choose to speak must address all of their arguments to all possible receivers. As we argue below, the equilibrium behavior that corresponds to this variant of the game furnishes the intuition for the main sequence we analyze. Finally, following the main analysis, we also consider two extensions. In the first extension, we analyze the incentives of the players to become speakers or listeners - thus addressing the question of the endogenous composition of the sets $\mathcal{S}$ and $\mathcal{R}$. (Note that with the exception of that extension, these sets of players are assumed fixed.) In our second extension, we consider the game in which the speakers observe which of the listeners' argument dimensions are latent and make publicly observable choices of which arguments, if any, to make on each of the possible dimensions.

[6] Examples of issues which would most immediately fall into the domain of private politics include the desirability of obtaining abortions (holding constant the official policy); the work-force 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.

section unless otherwise noted.[7] Thus, the receiver i learns r_i^j if and only if $r_i^j = s^j$ and r_i^j was latent. Let l_i be the number of "1"s and m_i be the number of "0"s that i learns from the speaker.

An immediate implication of the fact that the receivers make no inferences from "null" observations is that they treat these newly "learned" arguments as though they were, effectively, randomly revealed. Thus the receiver's posterior beliefs are characterized by

$$p(\theta_i|l+l', m+m') = \frac{\theta^{l+l'}(1-\theta)^{m+m'}}{\int_0^1 \hat{\theta}^{l+l'}(1-\hat{\theta})^{m+m'}d\hat{\theta}}. \tag{3}$$

Although the receiver does not comprehend fully what can be deduced from her observations, she does understand the stochastic nature of the process generating (l, m) to the extent that she can correctly assess the probability of obtaining a given outcome (l, m) given the speech σ and her prior beliefs about her type. Thus,

$$\Pr(l, m|\sigma, l', m', n) = \int_0^1 p(\theta|l', m') \Pr(l, m|\theta, \sigma, l', m', n)d\theta, \tag{4}$$

where

$$\begin{aligned}&\Pr(l, m|\theta, \sigma, l', m', n)\\ &= \tbinom{n-l'-m'}{l}\tbinom{n-l'-m'-l}{m}(\theta\sigma)^l((1-\theta)(1-\sigma))^m((1-\theta)\sigma+\theta(1-\sigma))^{n-l'-m'-l-m}.\end{aligned} \tag{5}$$

Let $Q \subseteq \mathcal{S}$ be the set of potential speakers who choose to speak, and let $g_i \in Q$ be the chosen speaker of $i \in \mathcal{R}$. An equilibrium strategy profile is a collection of sets $\langle Q^*, \{\sigma_i^*\}_{i\in Q}, \{g_i^*\}_{i\in\mathcal{R}}, \{\pi_i^*\}_{i\in\mathcal{R}}\rangle$, that characterizes sequentially rational choices for each player at each stage of the game, given their beliefs at the time of choice and the strategies of other players. The formal conditions that define such choices are given in the Appendix.

Although π_i^* is a function of the distribution of θ_i, we will suppress the argument. Substituting receiver i's optimal policy choice $\pi_i^* = E[\theta_i|p(\theta|l_i' + l_i, m_i' + m_i]$ into her expected utility, we obtain her perceived expected utility given her posterior beliefs

$$E[u_i(\pi_i^*|p(\theta|l_i + l_i', m_i + m_i'))] = -\mathrm{Var}(\theta_i|p(\theta|l_i + l_i', m_i + m_i'). \tag{6}$$

Thus, receiver i's (indirect) expected utility from hearing speech σ, given l_i', m_i', and n, is

$$U_i(\sigma, l_i', m_i', n) = -\sum_{l=0}^{n-l_i'-m_i'}\sum_{m=0}^{n-l_i'-m_i'-l} \Pr(l, m|\sigma, l_i', m_i', n)\mathrm{Var}(\theta_i|p(\theta|l + l_i', m + m_i')). \tag{7}$$

---

[7] We thus bracket all other possible deviations from logical omniscience. Our strategy here is similar to that of Rabin and Schrag [1999, p.38 fn2], 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 and of the behavioral patterns to which is gives rise.

Two baselines

In this section we present two baseline results that may be seen as offering intuitions for two central features of our model: (1) the deviation from the Bayesian updating on the part of the players and (2) allowing the players to join groups with particular speakers.

Our first result, then, concerns the behavior of a group of agents who update their beliefs using Bayes' Rule, in accordance with the standard model of rational choice under uncertainty. It establishes a striking baseline for comparison with the non-Bayesian listeners.

Proposition 1 *There is no speech in any group with Bayesian listeners.*

Proof See Appendix. ■

Because, given our interpretation of the model, the combination of Bayesian agency and the possibility of informative deliberation as self-discovery has no natural empirical referent, we do not offer a substantive interpretation of this result. Rather, we consider it a formal baseline that, together with the next proposition, clarifies the causal role of the cognitive limitation that we posit in (1). That said, Proposition 1 supplies another intuition for relaxing the assumption of Bayesian learning in a model of deliberation as self-discovery: given standard concave utilities, the expectation of Bayesian learning on the part of the listeners will preclude information transmission.

The formal intuition for this result is as follows. For every argument s_k^j that a speaker k makes, every listener i who has the matching latent argument $r_i^j = s_k^j$ is convinced by it and updates her beliefs about policy accordingly. Being Bayesian, every listener i for whom r_i^j is latent and $r_i^j \neq s_k^j$, must infer that she finds the opposite argument convincing– i.e. that her jth dimension argument is 1 if 0 was unconvincing and that it is 0 if 1 was unconvincing–and update her beliefs about policy accordingly. Because the listener is choosing her policy to maximize her expected utility given her beliefs, her policy choice (in the absence of further information) already strikes a balance between her possible true policy ideal points that takes into account the likelihood of her having each of them. Thus, the expected movement from learning another argument is 0. Because the speaker and the listener share the same ex ante beliefs about the listener's likely true policy ideal, the speaker has the same estimation of her listener's likely movement in response to additional information, but, because the speaker's utility exhibits diminishing marginal returns in the closeness of the listeners' policy choices, the speaker is harmed more by a listener's moving away than helped by her moving closer, ceteris paribus. This means that the speaker prefers not to speak at all.

With this result in mind, we assume in the remainder of the paper that agents learn in the manner characterized by (1). Before proceeding with our analysis of the main model, we characterize our second baseline result and a lemma that is instrumental in proving it and other results below. Say that a player is *biased* if her preferred policy position (or speech if indicated) is not $\frac{1}{2}$. We refer to the player as being left-biased if it is less than $\frac{1}{2}$ and right-biased if it is greater than $\frac{1}{2}$. The following result, then, obtains:

Lemma 1 *For any given biased speech σ, a listener i's policy position will, in expectation, move in the direction of σ's bias regardless of the relative sizes of l_i' and m_i'. If $\sigma = \frac{1}{2}$, the listener's policy position will, in expectation, remain unchanged.*

Proof See Appendix. ■

The intuition for this result is rather straightforward. If a speaker is more likely to produce right- than left-biased arguments (i.e., if $\sigma > \frac{1}{2}$), then a disproportionally high number of the listener's right-biased latent reasons will be activated. Thus, whatever the listener's initial bias, she will, in expectation, move to the right in response to a right-biased speaker.

We can now characterize our second baseline result. This result describes what happens when listeners update in the non-Bayesian fashion specified above and deliberate within a group that exhibits at least a minimal degree of ideological balance among its speakers (here we assume, in our first deviation from the main sequence of play, that players cannot form any "private conversation" groups). Call the group *balanced* if it includes at least one speaker who would make a left-biased speech and one speaker who would make right-biased speech *if speaking solo* to a given group of receivers. We consider two possibilities. In the first, the speakers must choose simultaneously whether or not to speak. In the other, they can choose to speak at any time, with the order of speech randomly determined for speakers who choose to speak at the same time, and the deliberation ends only when no speaker wishes to make any further (consequential) speech. The following proposition summarizes what happens in these cases:

Proposition 2 *In balanced groups,*

(1) if speech is simultaneous, then in equilibrium, all arguments are made and the speakers are worse off as a result;

(2) if speech is sequential and open-ended, then the Pareto optimal (for the speakers) outcome is equivalent to the outcome with fully Bayesian listeners - i.e., no speech takes place.

Proof See Appendix. ■

If speakers who with to move the listeners' policy choices in opposite directions must speak simultaneously, then each of them wishes to offer speech that is extreme enough to more than offset the effect of the other's speech. In equilibrium, then, they are driven to make the most extreme speeches possible, $\sigma = 0$ and $\sigma = 1$, ensuring that the listeners hear all arguments. In expectation, the listeners' positions will not change, but because the speakers' utilities exhibit diminishing marginal returns in the proximity of others' policy choices, the speakers are worse off. If, on the other hand, speakers may respond to the speech of others, and speech may continue until no one wishes to make further arguments, then the following strategies are an equilibrium: each speaker remains silent unless another speaker offers speech that causes a net movement away from her position; in that case, she responds with the "opposite" speech to counteract it. Because the speakers are worse off when the listeners hear both sets of arguments, no speech occurs.

Given the greater plausibility of sequential over simultaneous speech, Proposition 2 suggest that we should expect the same outcome in balanced groups as in the groups with Bayesian listeners even when the listeners are not Bayesian in the precise sense specified above. Proposition 2 may, moreover, be interpreted as suggesting that speakers have a preference for establishing separate forums - indeed, a dominant strategy to do so - implying the importance of considering how the listeners respond to the necessity of choosing among

speakers. That realization is one of the motives for the game-form we analyze in the next section.

Deliberation in biased groups

In this section we consider the main sequence of play (recall: each speaker i publicly announces a speech parameter σ, the listeners decide which speaker to join, hear the corresponding speech, and make their policy choices based on the updated beliefs). Our primary focus is on the listeners' choices of speakers and their consequences for the post-deliberative ideological composition of the society.

Before proceeding with the characterization of those choices, we draw attention to two key features of the speakers' optimal choices. First, those choices are well-defined, that is, there exists an optimal choice of σ_i for every speaker i, given the choices of other speakers and the strategies of the receivers. To see this, note, given (2) - (7), that speaker i's expected utility is continuous in σ_i for $\sigma_i \in [0, 1]$. Every function that is continuous on a closed interval has at least one maximum on that interval, and thus there exists a value of σ_i that maximizes the speaker's expected utility.

The second feature of the speakers' choices is that the ideological bias of optimal speech is monotonic in the bias of the speaker. More precisely, we have the following lemma:

Lemma 2 *Speaker i's optimal choice of σ_i is weakly increasing in $E[\theta_i]$.*

Proof See Appendix. ■

Thus, the greater the right (left) ideological bias of the speaker, the weakly greater the right (left) bias of her speech in equilibrium. Given the optimal choices of speech, what should we expect from the listeners? Our first answer is the following proposition:

Proposition 3 *(Biased Groups) A listener always prefers to join a group offering the most extreme speech σ in the direction of her own pre-deliberative bias, i.e., the most right-biased speech if $l' > m'$ and the most left-biased speech if $m' > l'$. When $l' = m'$ the listener prefers to join the group with the most extreme speech, regardless of the direction of the ideological bias.*

Proof See Appendix. ■

Proposition 3 shows that listeners prefer to hear arguments from the most extreme speakers whose biases are in the same direction as their own. This means, *inter alia*, that we should expect deliberative groups to consist of agents who are, in expectation, biased in the same ideological direction.

This result depends on the agents' failure of negative introspection in several ways. First, because they are not negatively introspective, they will learn only when they hear the argument that matches their own latent reason on the same dimension. Thus they wish to hear the speaker who is going to supply (in expectation) the most arguments consistent with their own pre-deliberative bias. For example, since each of a right-biased listener's latent reasons is more likely to be right-biased than left-biased, she obtains the most "matches" in expectation from a speaker who offers the right-biased argument on every dimension. Second, because they are not negatively introspective, they are incapable of realizing that their

post-deliberative active arguments are a result of selection bias. As a result, they feel more certain in their post-deliberative policy choice (i.e., they believe that their true ideal point is more likely to be close to their expectation of it) when their active arguments are more biased toward an extreme, and thus their expected utility as they perceive it is higher when they have been exposed to a more biased source. This perception makes hearing the more extreme speaker still more desirable in the eyes of the listener.[8]

The conjunction of Proposition 3 and Lemma 1 yields the following corollary:

Corollary 1 *(Social Polarization) If there is a speaker with $\sigma > \frac{1}{2}$ ($\sigma < \frac{1}{2}$), then, in expectation, listeners for whom $l' > m'$ ($m' > l'$) will choose post-deliberative policies that are, on average, more right-(left-) biased than their pre-deliberative positions.*

This corollary may be seen as identifying the ultimate effects of individual decisions regarding group membership. Because, by Proposition 3, listeners will prefer to join the groups with speech that is most biased in the direction of their own pre-deliberative bias, free group choice, given Lemma 1, must, in the presence of both left-of-center and right-of-center speech, lead to greater post-deliberative ideological polarization.

Our next result describes another interesting feature of the listeners' preferences that further underscores the expectation of post-deliberative ideological bias when citizens exercise the freedom of listening to the speakers of their choice. The following proposition shows that listeners may be diverse not only in the ideological direction of the deliberative groups to which they may, all things considered, prefer to belong, but also in their interest in hearing arguments of particular ideological stripes when doing so has no opportunity costs. Perhaps surprisingly, it shows that while some listeners prefer to hear speech no matter what ideological bias it displays (though if they had to make a choice, they would still prefer, by Proposition 3, to hear the speech that reinforces their own bias), others may prefer less to more information:

Proposition 4 *For any number of argument dimensions $n \geq 3$, one of the following statements is true for every, and each statement is true for at least one, possible type of listener (l', m'):*

(1) the listener prefers hearing nothing to hearing a speaker whose speech is sufficiently left-biased;

(2) the listener prefers hearing any speaker's speech to hearing nothing;

(3) the listener prefers hearing nothing to hearing a speaker whose speech is sufficiently right-biased.

Proof See Appendix. ■

The core intuition for this result is as follows. Because the listener treats her active reasons as a random sample of all her reasons, she perceives less uncertainty about her true

[8]Note that, because of the nature of agents' learning, repeated choice of the group is always going to be in the direction of the prior bias: agents never realize that the continuing presence of the latent argument dimensions means that they are likely to be more convinced by the arguments made by an oppositely biased speaker.

policy ideal, θ, when her set of active reasons corresponds to a more extreme expected policy ideal. Thus, her own evaluation of her expected indirect utility is greater, ceteris paribus, for more extreme expected policy ideals. Because she is not negatively introspective, she fails to realize that her post-deliberative active arguments display selection bias, and thus treats them as a random sample, as well. For some listeners, hearing speech from the opposing extreme is likely to increase the variance of the listener's beliefs about her true ideal policy, thus reducing her perceived expected indirect utility. Ex ante, such a listener would prefer not to hear such speech.

Apart from describing an interesting epistemic consequence of the cognitive agency in our model, Proposition 4 may be interpreted as providing an epistemic explanation for the pressures to conform within a deliberative group. Although we are not modeling this possibility, it is natural to ask whether and to what extent we should expect to see dissenting speech within groups. It is clear from Proposition 2 that, within each group, the speaker has a preference against others engaging in such (uncoordinated or unsanctioned) speech. Proposition 4 goes a step further - it points to the presence of listeners who are interested in suppressing a dissenting speech in their group out of concern for the quality of their own judgments and quite apart from the effects that that speech may have on others. Although, as Proposition 4 shows, there may also be listeners who would welcome the dissenting speech within the group, those listeners will also prefer the group with the single most biased speaker to one with the multiple potential conflicting speakers because, by Proposition 2, the possibility of dissenting speech will, in fact, mean less or no speech in equilibrium.

Having developed the expectation of the direction of post-deliberative movement, it is natural to ask how uniform in size we should expect the ideological movement within groups to be. The answer will certainly depend on the composition of the group. In particular, it will be affected by the differences in the extent of knowledge on the part of the group members - in the context of our model, on how many active reasons members of the group know relative to one another. We can, however, give a clean answer ceteris paribus - assuming away the variation in the number of known arguments. As our next proposition shows, the magnitude of the expected ideological shift is non-monotonic in listeners' pre-deliberative ideological position:

Proposition 5 *Holding constant the number of latent dimensions on which the speech is heard,*

(1) if $\sigma > \frac{1}{2}$, the expected rightward shift is increasing in l' for left-biased listeners and decreasing in l' for right-biased listeners;

(2) if $\sigma < \frac{1}{2}$, the expected leftward shift is increasing in m' for right-biased listeners and decreasing in m' for left-biased listeners.

Proof See Appendix. ■

The number of i's left or right active arguments determines the expected effect of speech on her policy choice through two channels. First, in that they reflect information about i's true policy ideal, θ_i, they also indicate how likely i is to learn from speech σ: the more right-biased (l'_i, m'_i), the more likely i is to recognize arguments when she listens to right-biased speech, $\sigma > \frac{1}{2}$. Second, they determine how much impact an additional active argument has on i's policy choice, $E[\theta_i]$: if more initial active arguments are right-biased then the

impact of an additional right-biased active argument is smaller. One consequence of this is that i's recognition of an additional argument in the same direction as her initial bias has a smaller effect on her policy choice than does her recognition of an additional argument in the opposite direction. In the presence of biased speech (which the listener essentially treats as unbiased in updating her beliefs), this asymmetry produces the nonmonotonic response described in the result.

If the effectiveness of speech is measured by the expected difference between pre- and post- deliberative positions, then this proposition implies that a speaker seeking to be most effective will be happiest with an audience composed of the most moderate listeners. From the standpoint of maximizing her own utility, the right-biased speaker would prefer the left-biased audience. Alas, as Proposition 3 shows, those listeners would prefer a speaker with the opposite bias. The closer a listener is to the speaker, while sharing the direction of bias, the less effective the speaker is likely to be with respect to that listener.

Endogenous Speakers

The analysis in the previous section has assumed fixed sets of potential speakers and potential listeners. In our first extension, we consider the incentives individuals face for becoming speakers or listeners. Consider a game in which every player i chooses to send or to receive; if i chooses to send, she must then choose $\sigma_i \in [0,1]$; if i chooses to receive, she must then choose which one of the speaking agents to hear.[9] Following these choices, each player i who chose to receive learns (l_i, m_i) and updates her beliefs. Each player then chooses her policy π_i and realizes her payoff based on the sum of squared differences between her ideal point and each player's policy choice. As before, we require that the equilibrium strategies satisfy sequential rationality, i.e., that each player's chosen action be optimal given her beliefs at the time of action, and that beliefs be updated in the manner described by (3).

In such a game, the players must weigh the benefits of speaking (of potentially influencing others' policy choices) against the benefits of listening (of changing one's own policy choice) where final payoffs depend on the policy choices of all agents. Lemma 2 implies that optimal speeches by more ideologically extreme speakers must also be more extreme. But who chooses to speak? Should we expect to see the more extreme speakers, or is there a tendency toward moderation among the speakers? Our next result shows that we should expect the former, and not the latter.

Proposition 6 *In equilibrium, the players with the most extreme left- and right- biased policy positions choose to speak, ceteris paribus.*

Proof See Appendix. ■

Thus, not only should we expect biased speech, but we should, in the long run, expect the most extreme speakers to be at least as extreme as the listeners who are biased in the same direction. The intuition here is as follows. The diminishing marginal returns to

[9] For interpretive reasons, it is natural to suppose that agents can make arguments only on those dimensions on which their own reasons are active. This assumption has no independent technical bite, and thus the results are the same whether or not it is imposed.

the proximity of others' policy choices guarantees that a more extreme right-biased agent will benefit more from the right-ward movement of listener's policy choices than will a less extreme right-biased agent, and thus will benefit more from speaking. The fact that more extreme agents are less responsive to speech (established in Proposition 5) implies that they benefit less from listening. Together these facts yield the result.

Given Lemma 1, Proposition 6 implies that if there is any speech in equilibrium, we should expect to observe social polarization. There is another noteworthy consequence as well. As the proof of Proposition 6 shows, the most ideologically biased agents have the highest net benefits from speaking. This fact is a further reinforcement of the intuition, already noted in the context of Proposition 4, that we should expect speech in deliberating groups to be dominated by the most extreme members - i.e., that the single-speaker-group model we analyze in the present paper is a plausible approximation to the richer environment with multiple possible speakers within each group.

Targeted Speech

Our second extension concerns a variant of the main game in which both the speakers and the listeners have additional information about each other. In particular, suppose that speakers observe which of the listeners' argument dimensions are latent and make publicly observable choices of which arguments, if any, to make on each of the possible dimensions. The listeners are, thus, making their choices of which speaker or group to join knowing both on which dimensions a given speaker will speak and what kind of argument they will be hearing on each relevant dimension, "0" or "1."

First observe that Propositions 1 and 2 are unaffected by this change (recall that the latter is proven inductively for communication on k latent dimensions). Thus, we should still expect speakers to have a preference for creating specialized exclusive forums. The expected response of the listeners' policy choices to speech is still movement in the direction of the speech's bias; a listener who hears predominantly "1"s on her latent argument dimensions moves rightward in expectation. The key outstanding question is whether we should also expect an equivalent of Proposition 3.

Let Σ_i be the set of dimensions on which speaker i makes arguments, $\Sigma_i = \Sigma_i^0 \cup \Sigma_i^1$, where Σ_i^0 is the set of all "0" arguments, and Σ_i^1 the set of all "1"arguments in her speech, i.e., $t \in \Sigma_i^1$ if and only if $s_i^t = 1$ and $t \in \Sigma_i^0$ if and only if $s_i^t = 0$. We say that Σ_i is *full coverage* if i's arguments are adduced on every dimension $(1, ..., n)$, that is $|\Sigma_i^0 \cup \Sigma_i^1| = n$. (Note that in the model analyzed above, the mechanism generating arguments is such that Σ_i is full coverage for each $i \in \mathcal{S}$.) Correspondingly, we say that Σ_i is *not full coverage* if $|\Sigma_i^0 \cup \Sigma_i^1| < n$. Recall, finally, that g_j is j's choice of which speaker to hear. We can now state the following result:

Proposition 7 *Let i be the most extreme right- (left-)biased potential speaker. If $\frac{1}{2} < E[\theta_h] < E[\theta_i]$ ($\frac{1}{2} > E[\theta_h] > E[\theta_i]$), then in equilibrium there is $j \in \mathcal{R}$ s.t. $E[\theta_j] > \frac{1}{2}$ ($E[\theta_j] < \frac{1}{2}$) and $g_j = h$ if and only if*

(1) $E[\theta_i]$ is sufficiently moderate;

(2) for each listener j such that $g_j^* = h$, there is a listener k such that $g_k^* = i$ and $\mathcal{L}'_j \cap \mathcal{L}'_k \neq \emptyset$;

(3) if $\Sigma_{i,h}$ are full coverage, then h's speaking is weakly Pareto optimal for i and h;

(4) if $\Sigma_{i,h}$ are not full coverage, then either h's speaking is weakly Pareto optimal for i and h or h's audience is small compared to i's audience.

Proof See Appendix. ■

Proposition 7 shows what can be expected in equilibrium when the speakers and listeners have more information about each other. Its notable feature is the possibility of less extreme speakers in equilibrium. However, the circumstances under which this possibility is realized are relatively special. Note in particular the conjunction of conditions (1) and (2). If the more extreme speaker is relatively moderate then an extreme speech from her may cause some members of her audience to become too extreme from her standpoint. At the same time, she may want to move other, more moderate, listeners by a substantially greater distance in her (more extreme) direction, and so may prefer to have them hear a more extreme speech. There may, thus, be a conflict between the kinds of speech she would like to offer to these different types of listeners. But this conflict can occur only if there is sufficient overlap between those prospective listeners' latent argument dimensions - since, otherwise, the single speech would be equivalent to different speeches targeting different segments of the audience. The possibility of such a conflict creates an opening for the less extreme speakers. In particular, the more extreme speaker may offer speech more suitable to the more extreme listeners, anticipating that the less extreme speaker will attract the still less extreme listeners whom she may prefer to move closer to her by making more extreme arguments. If Σ is full coverage, this must, by Proposition 3, mean that *on the dimensions that are latent for those listeners*, the less extreme speaker's speech is more extreme than that of the speaker with more extreme policy preferences. In this case, then, the more extreme speaker is better off, but so also is the less extreme speaker, both because she moves the less extreme listeners closer to her own position and because the more extreme listeners do not move as far to the right (and away from her) as they otherwise would have if the more extreme speaker were speaking solo. Somewhat paradoxically, the possibility of a more moderate speaker's speech can, thus, moderate the post-deliberative positions of the most extreme listeners even though on the dimensions on which her speech reaches her audience it is more extreme than that of the more extreme speaker.

When Σ is not full coverage, the less extreme speaker may be able to attract an audience by offering more arguments than the more extreme speaker even when the latter's speech is more extreme (has a higher proportion of "1" for the right-biased speakers or "0" for the left-biased ones). In such cases, the more extreme speaker can be worse off and will prefer to compete for the audience by offering more arguments. In fact, the argument of the proof shows that she will prefer to do so to the point that reduces the less extreme speaker's audience to a size that is smaller than her own. The intuition is that the combination of speakers' concave utilities and Proposition 5 means that the more extreme speaker will prefer the marginal "overshooting" by the more extreme listeners to the marginal "undershooting" by the less extreme ones. Thus, when she does not benefit from the fact that the less extreme speaker draws audience, her own optimal choices can be expected to substantially reduce that speaker's influence.

The Presence of Bayesians

Our last extension concerns the possibility that population may contain standard Bayesian agents alongside the agents with the non-negatively introspective agents whose behavior has been our focus in the preceding. Are the substantive results robust to this possibility? Should we still expect to see group polarization in this case?

First, observe that Bayesian speakers will choose the same speech as their non-Bayesian counterparts if they have the same expectations of the listeners' responses. Thus, the results will not be affected by assuming that the speakers, who may be thought of as corresponding to political elites, are more capable in this respect than the listeners.

Second, consider the possibility that speakers and listeners are equally likely to be Bayesian. From Proposition 1, Bayesian speakers will not want to speak to Bayesian listeners because they anticipate that the expected change in the Bayesian listeners' policy choices is zero, and because the diminishing marginal returns of policy proximity ensures that the speaker's loss from one listener's moving away is greater than the benefit from another's moving closer. Thus, if they do speak, it must be that there are enough non-Bayesian listeners who will, in expectation, move closer to them to make speaking more desirable than being silent. Because non-Bayesians seek the most extreme speakers who share their bias (whereas Bayesians are indifferent between all speakers and draw the same conclusions from any s), Bayesian listeners' post-deliberative policy choices will, on average, be more extreme in the direction of their prior biases, and non-Bayesians' post-deliberative positions will, on average, be the same, resulting in a more extreme average post-deliberative position for the group.

Non-Bayesian speakers do not recognize that Bayesian listeners update their beliefs in response to unpersuasive arguments, and thus such speakers' behavior will be like that described in the main results. Again, this speech will not affect the average position of the Bayesian listeners, but will cause the non-Bayesian listeners to become more polarized. Thus, although the extent of social polarization would be less in such a mixed population, it would still occur.

IV. Discussion

The model in this paper captures 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 her message and the analytical structure of the listener's current beliefs. 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 that is strongly consistent with empirical evidence, but also in a way that preserves the optimality property of individual choices, given beliefs.

We show, *inter alia*, that non-negatively introspective agents prefer to seek out the most extreme speakers whose bias re-enforces their own, and that these decisions on deliberative group membership give rise, in equilibrium, to the phenomenon of post-deliberative group polarization. It is worthwhile to note that, from a theoretical standpoint, this result is

a consequence of the combination of two factors that distinguish the present model: the particular aspects of the cognitive agency being analyzed and the fact that, though the arguments we are modeling are provable (removing all uncertainty directly), they do not have common veridicality for the agents. Holding all else fixed, doing away with the agents' cognitive limitation yields Proposition 1, which makes the group polarization effect, and thus the social polarization effect, moot. On the other hand, retaining our model of the cognitive agency but assuming common veridicality would have the effect of making everyone's pre-deliberative arguments equally informative, enabling the agents to use that information to converge both on the same pre-deliberative policy beliefs and on the same preferences over deliberating groups. (Indeed, speakers themselves would no longer have a preference for seeking exclusive forums.) Here, too, the sorting of speakers and listeners that gives rise to social polarization would not be expected.

Although our explanation of social polarization is "informational," it relies on individual choices regarding group membership that are, in turn, a consequence of particular aspects of cognition. Because individuals learn and conceive of themselves as learning in the fashion we characterize, they seek out speakers that provide the strongest possible reinforcement of their prior biases. It is such group membership choices that, given the listeners' cognition, lead to social polarization. In short, a specific wide-spread feature of individual cognition may be seen as responsible for both which deliberative groups individuals prefer to join and, along with the effects of those choices, for what consequences of group deliberation we should expect.

V. Appendix

Definition of the Equilibrium

The equilibrium of the game is defined by the beliefs described in the main text and a strategy profile $\langle Q^*, \{\sigma_i^*\}_{i\in Q}, \{g_i^*\}_{i\in\mathcal{R}}, \{\pi_i^*\}_{i\in\mathcal{R}}\rangle$ which satisfies each of the following conditions:

1. $i \in Q^*$ iff

$$-\int_0^1 p(\theta|l_i', m_i')\left(\sum_{j\in\mathcal{R}}\sum_{l=0}^{n-l_j'-m_j'}\sum_{m=0}^{n-l_j'-m_j'-l}\Pr(l,m|g_j^*(\{\sigma_k^*\}_{k\in Q^*\setminus\{i\}}),\sigma_{g_j^*}^*,l_j',m_j',n)(\pi_j^*(\cdot)-\theta)^2\right)d\theta$$
$$<-\int_0^1 p(\theta|l_i', m_i')\left(\sum_{j\in\mathcal{R}}\sum_{l=0}^{n-l_j'-m_j'}\sum_{m=0}^{n-l_j'-m_j'-l}\Pr(l,m|g_j^*(\{\sigma_k^*\}_{k\in Q^*\cup\{i\}}),\sigma_{g_j^*}^*,l_j',m_j',n)(\pi_j^*(\cdot)-\theta)^2\right)d\theta;$$

2. $\forall i \in Q^*$, $\sigma_i^* \in \arg\max_{\sigma_{i\in[0,1]}}$

$$\left(-\int_0^1 p(\theta|l_i', m_i')\left(\sum_{j\in\mathcal{R}}\sum_{l=0}^{n-l_j'-m_j'}\sum_{m=0}^{n-l_j'-m_j'-l}\Pr(l,m|g_j^*(\{\sigma_i\}\cup\{\sigma_k^*\}_{k\in Q^*}),\sigma_{g_j^*}^*,l_j',m_j',n)(\pi_j^*(\cdot)-\theta)^2\right)d\theta\right);$$

3. $\forall i \in \mathcal{R}$,

$$g_i^* \in \arg\max_{g_i\in Q^*}\left(-\sum_{l=0}^{n-l_i'-m_i'}\sum_{m=0}^{n-l_i'-m_i'-l}\Pr(l,m|g,\sigma_{g_i^*}^*,l_i',m_i',n)\int_0^1 p(\theta|l,m,l_i',m_i')(\pi_i^*(\cdot)-\theta)^2)d\theta\right);$$

4. $\forall i \in \mathcal{R}$, $\pi_i^* \in \arg\max_{\pi_i\in[0,1]}\left(-\int_0^1 p(\theta|l_i,m_i,l_i',m_i')(\pi_i^*(\cdot)-\theta)^2)d\theta\right)$.

Proofs of Formal Results

Proposition 1

Proof We show, by induction, that a speaker is, in expectation, worse off if she speaks to a Bayesian audience. Because the listener is Bayesian and only two arguments are possible, $\{0,1\}$, she learns her argument on every dimension if the speaker speaks. First, we show that the speaker is worse off if the listener learns one dimension than if she learns none. We then show that the speaker is worse off if the listener learns $k+1$ dimensions of r than if she learns k.

Given that $\pi_j^* = E[\theta_j | p(\theta|\cdot)]$, listener j chooses

$$\pi_j^* = \int_0^1 \theta p(\theta|l_j', m_j')d\theta = \frac{\int_0^1 \theta^{l_j'+1}(1-\theta)^{m_j'}d\theta}{\int_0^1 \theta^{l_j'}(1-\theta)^{m_j'}d\theta} = \frac{l_j'+1}{l_j'+m_j'+2} \tag{8}$$

if the speaker is silent. Suppose listener j learns one dimension (i.e., $n = l_j' + m_j' + 1$). She chooses $\pi_j^* = \frac{l_j'+2}{l_j'+m_j'+3}$ if she infers argument "1," and $\pi_j^* = \frac{l_j'+1}{l_j'+m_j'+3}$ if she infers argument "0." The probability of her learning "1" is θ_j, so the expected probability of "1" is (8). Her expected probability of "0" is

$$E[1-\theta_j | p(\theta|l_j', m_j')] = \frac{\int_0^1 \theta^{l_j'}(1-\theta)^{m_j'+1}d\theta}{\int_0^1 \theta^{l_j'}(1-\theta)^{m_j'}d\theta} = \frac{m_j'+1}{l_j'+m_j'+2}.$$

Let l_i', m_i' be speaker i's known numbers of ones and zeroes, respectively. i's expected utility from j's choice if she is silent is greater than if she speaks iff

$$\begin{aligned} & -\int_0^1 p(\theta|l_i', m_i')\left(\frac{l_j'+1}{l_j'+m_j'+2} - \theta\right)^2 d\theta \\ > \; & -\left(\frac{l_j'+1}{l_j'+m_j'+2}\right)\int_0^1 p(\theta|l_i', m_i')\left(\frac{l_j'+2}{l_j'+m_j'+3} - \theta\right)^2 d\theta \\ & -\left(\frac{m_j'+1}{l_j'+m_j'+2}\right)\int_0^1 p(\theta|l_i', m_i')\left(\frac{l_j'+1}{l_j'+m_j'+3} - \theta\right)^2 d\theta. \end{aligned}$$

Taking expectations, we obtain an equivalent inequality

$$\begin{aligned} & -\left(\frac{l_j'+1}{l_j'+m_j'+2}\right) + 2\left(\frac{l_j'+1}{l_j'+m_j'+2}\right)E[\theta_i] - E[\theta_i^2] \\ > & -\left(\frac{l_j'+1}{l_j'+m_j'+2}\right)\left[\left(\frac{l_j'+2}{l_j'+m_j'+3}\right)^2 - 2\left(\frac{l_j'+2}{l_j+m_j+3}\right)E[\theta_i] + E[\theta_i^2]\right] \\ & -\left(\frac{m_j'+1}{l_j'+m_j'+2}\right)\left[\left(\frac{l_j'+1}{l_j'+m_j'+3}\right)^2 - 2\left(\frac{l_j'+1}{l_j+m_j+3}\right)E[\theta_i] + E[\theta_i^2]\right], \end{aligned} \tag{9}$$

which is always true.

Suppose now that the speaker prefers the listener learning nothing to learning k dimensions. Define l'' and m'' such that $l'' \geq l'$, $m'' \geq m'$, and $(l''-l')+(m''-m') = k$. Substituting l'' for l' and m'' for m' in equation (9) yields the condition for the speaker preferring the listener's learning only k dimensions to her learning $(k+1)$. Because this argument is independent of the group composition and the speaker's type, it must hold for any group of Bayesian listeners. ■

Lemma 1

Proof Solving for π_i^* using (3), we obtain $\pi_i^* = \frac{l+l'+1}{l+l'+m+m'+2}$. Thus, $\pi_i^* > \pi_i^{*\prime}$ if and only if $\frac{l+l'+1}{l+l'+m+m'+2} > \frac{l'+1}{l'+m'+2}$, i.e., if and only if $l > \frac{l'+1}{m'+1}m$. (l, m) satisfy this condition in expectation if and only if

$$\sigma E[\theta_i|p(\theta|l', m')] > \frac{l'+1}{m'+1}(1-\sigma)(1-E[\theta_i|p(\theta|l', m'))],$$

which is equivalent to

$$\sigma\frac{l'+1}{l'+m'+2} > \frac{l'+1}{m'+1}(1-\sigma)(1-\frac{l'+1}{l'+m'+2}),$$

which simplifies to $\sigma > \frac{1}{2}$. The argument is symmetric for $\pi_i^* < \pi_i^{*\prime}$. ■

Lemma 2

Proof For every potential pair l, m, there is a post-deliberative probability density function $p(\theta|l+l', m+m')$ that induces receiver j's $\pi_j^* = E[\theta_j|p(\theta|l+l'_j, m+m'_j)]$. A probability distribution over possible values $E[\theta_j|p(\theta|l+l'_j, m+m'_j)]$ is given by (4). Hence, the component of speaker i's utility determined by any given receiver's behavior is, in expectation,

$$\begin{aligned}
& -\sum_{l=0}^{n-l'_j-m'_j}\sum_{m=0}^{n-l'_j-m'_j-l}\Pr(l,m|\sigma,l'_j,m'_j,n)\int_0^1 p(\theta|l'_i,m'_i)(E[\theta_j|p(\theta|l+l'_j,m+m'_j)]-\theta)^2 d\theta \\
= \; & -E[\theta_i^2|p(\theta|l'_i,m'_i)] - \sum_{l=0}^{n-l'_j-m'_j}\sum_{m=0}^{n-l'_j-m'_j-l}\Pr(l,m|\sigma,l'_j,m'_j,n)(E[\theta_j|p(\theta|l+l'_j,m+m'_j)])^2 \\
& +2E[\theta_i|p(\theta|l'_i,m'_i)]\sum_{l=0}^{n-l'_j-m'_j}\sum_{m=0}^{n-l'_j-m'_j-l}\Pr(l,m|\sigma,l'_j,m'_j,n)E[\theta_j|p(\theta|l+l'_j,m+m'_j)]. \quad (10)
\end{aligned}$$

Let $E[\theta_h|\cdot] > E[\theta_i|\cdot]$. Then, if σ_h and σ_i are h and i's optimal choices of speech, respectively, it follows that h's indirect expected utility from σ_h is at least as great as her expected utility from σ_i, and i's indirect expected utility from σ_i is at least as great as her expected

utility from σ_h. From (10), we obtain

$$-E[\theta_h^2|\cdot] - \sum_{l=0}^{n-l'_j-m'_j} \sum_{m=0}^{n-l'_j-m'_j-l} \Pr(l,m|\sigma_h,\cdot)E[\theta_j|\cdot](E[\theta_j|\cdot] - 2E[\theta_h|\cdot])$$
$$\geq -E[\theta_h^2|\cdot] - \sum_{l=0}^{n-l'_j-m'_j} \sum_{m=0}^{n-l'_j-m'_j-l} \Pr(l,m|\sigma_i,\cdot)E[\theta_j|\cdot](E[\theta_j|\cdot] - 2E[\theta_h|\cdot]).$$

and

$$-E[\theta_i^2|\cdot] - \sum_{l=0}^{n-l'_j-m'_j} \sum_{m=0}^{n-l'_j-m'_j-l} \Pr(l,m|\sigma_i,\cdot)E[\theta_j|\cdot](E[\theta_j|\cdot] - 2E[\theta_i|\cdot])$$
$$\geq -E[\theta_i^2|\cdot] - \sum_{l=0}^{n-l'_j-m'_j} \sum_{m=0}^{n-l'_j-m'_j-l} \Pr(l,m|\sigma_h,\cdot)E[\theta_j|\cdot](E[\theta_j|\cdot] - 2E[\theta_i|\cdot]).$$

Subtracting the bottom inequality from the top and simplifying, we obtain a true inequality

$$\sum_{l=0}^{n-l'_j-m'_j} \sum_{m=0}^{n-l'_j-m'_j-l} \Pr(l,m|\sigma,\cdot)E[\theta_j|\cdot](E[\theta_i|\cdot] - E[\theta_h|\cdot]) \tag{11}$$
$$\geq \sum_{l=0}^{n-l'_j-m'_j} \sum_{m=0}^{n-l'_j-m'_j-l} \Pr(l,m|\sigma_h,\cdot)E[\theta_j|\cdot](E[\theta_i|\cdot] - E[\theta_h|\cdot]).$$

Given that $E[\theta_i|\cdot] < E[\theta_h|\cdot]$, the inequality (11) holds only if σ_h results in more probability mass on those outcomes (l,m) for which $E[\theta_j|p(\theta|l+l'_j, m+m'_j)]$ is higher, i.e., $\sigma_h \geq \sigma_i$. ■

Proposition 2

Proof Consider two potential speakers, 1 and 2, such that 1's optimal speech when she alone speaks is $\hat{\sigma}_1 < \frac{1}{2}$ and 2's optimal speech when she alone speaks is $\hat{\sigma}_2 > \frac{1}{2}$. From Lemma 1, in expectation, the receivers will move left in response to $\hat{\sigma}_1$ and right in response to $\hat{\sigma}_2$. The optimality of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ and the given strategy space implies that 1 wants the receivers to move left and 2 wants them to move right. If 1 is offering speech $\hat{\sigma}_1$, then 2, in order to achieve her desired net move of the receivers to the right, must offer $\sigma'_2 > \hat{\sigma}_2$. Similarly, in the presence of 2's speech, 1 must prefer a more extreme leftist speech $\sigma'_1 < \hat{\sigma}_1$. But then 2 must prefer $\sigma_2 > \sigma'_2$, etc. Thus, in equilibrium, $\sigma_1^* = 0$ and $\sigma_2^* = 1$. Every receiver $i \in \mathcal{R}$ learns r^i, and by the same argument made in Proposition 1, every speaker obtains lower utility in expectation from the members' policy choices when they are informed than she does when they are uninformed.

(2) Because the making of arguments ends endogenously, i.e., there is no designated last speaker or last period, the following speaker strategy subprofile, combined with π^*, is a subgame perfect equilibrium: for every speaker i, "be silent if others are silent; speak if some speaker k speaks and the change in $\{\pi_j\}_{j\in\mathcal{R}}$ in response to s_k reduces your utility." Given that listeners choose π optimally in expectation and that those who updated in response to

s_k moved, on average, away from i, those listeners who will update in response to s_i will move, in expectation, toward i, increasing i's utility. Thus, speaking is a better response to a harmful speech than remaining silent. However, being silent is a better response to silence given the other players' strategies because, as shown in the proof of Proposition 1, every speaker's utility decreases in expectation if listeners resolve their uncertainty. ■

Proposition 3

Proof We proceed in three steps.

Step 1: The expected utility of a receiver, given l, l', m, and m', is increasing faster in $(l + l')$ than in $(m + m')$ if and only if $(l + l') > (m + m')$.

In effect, a listener i updates as if $l + m$ dimensions have been randomly revealed to her. Given the posteriors (3), i chooses $\pi_i^* = E[\theta_i | p(\theta | l + l'_i, m + m'_i)]$, which induces expected utility $E[u_i(\pi_i^* | p(\theta | l + l'_i, m + m'_i))] = -\text{Var}(\theta_i | p(\theta | l + l'_i, m + m'_i))$.

$$\begin{aligned}
& \text{Var}(\theta_i | p(\theta | l + l'_i, m + m'_i)) \\
= \; & E[\theta_i^2 | p(\theta | l + l'_i, m + m'_i)] - (E[\theta_i | p(\theta | l + l'_i, m + m'_i)])^2 \\
= \; & \frac{\int_0^1 \theta^{l+l'+2}(1-\theta)^{m+m'} d\theta}{\int_0^1 \theta^{l+l'}(1-\theta)^{m+m'} d\theta} - \left(\frac{\int_0^1 \theta^{l+l'+1}(1-\theta)^{m+m'} d\theta}{\int_0^1 \theta^{l+l'}(1-\theta)^{m+m'} d\theta} \right)^2 \\
= \; & \frac{\Gamma(l + l'_i + 3)\Gamma(m + m'_i + l + l'_i + 2)}{\Gamma(l + l'_i + 1)\Gamma(m + m'_i + l + l'_i + 4)} - \left(\frac{\Gamma(l + l'_i + 2)\Gamma(m + m'_i + l + l'_i + 2)}{\Gamma(l + l'_i + 1)\Gamma(m + m'_i + l + l'_i + 3)} \right)^2,
\end{aligned}$$

where $\Gamma(\cdot)$ is Euler's gamma function. Substituting equivalent expressions and simplifying, we obtain

$$\begin{aligned}
& E[u_i(\pi_i^* | p(\theta | l + l'_i, m + m'_i))] \\
= \; & -\frac{(l + l'_i + 2)!(m + m'_i + l + l'_i + 1)!}{(l + l'_i)!(m + m'_i + l + l'_i + 3)!} + \left(\frac{(l + l'_i + 1)!(m + m'_i + l + l'_i + 1)!}{(l + l'_i)!(m + m'_i + l + l'_i + 2)!} \right)^2 \\
= \; & -\frac{(l + l'_i + 1)(m + m'_i + 1)}{(m + m'_i + l + l'_i + 2)^2(m + m'_i + l + l'_i + 3)}. \qquad (12)
\end{aligned}$$

Differentiating with respect to l and m, we get

$$\begin{aligned}
& \frac{\partial E[u_i(\pi_i^* | p(\theta | l + l'_i, m + m'_i))]}{\partial l} \\
= \; & \frac{(m + m'_i + 1)((2 - 2m - 2m'_i) + 6(l + l'_i) + 2(l + l'_i)^2 + (l + l'_i)(m + m'_i) - (m + m'_i)^2)}{(l + l'_i + m + m'_i + 2)^3(l + l'_i + m + m'_i + 3)^2} \qquad (13) \\
& \frac{\partial E[u_i(\pi_i^* | p(\theta | l + l'_i, m + m'_i))]}{\partial m} \\
= \; & \frac{(l + l'_i + 1)((2 - 2l - 2l'_i) + 6(m + m'_i) + 2(m + m'_i)^2 + (l + l'_i)(m + m'_i) - (l + l'_i)^2)}{(l + l'_i + m + m'_i + 2)^3(l + l'_i + m + m'_i + 3)^2}.
\end{aligned}$$

It follows that

$$\frac{\partial E[u_i(\pi_i^*|p(\theta|l+l_i', m+m_i'))]}{\partial l} > \frac{\partial E[u_i(\pi_i^*|p(\theta|l+l_i', m+m_i'))]}{\partial m}$$
$$\text{iff} \qquad -(l+l_i'-(m+m_i'))(l+l_i'+m+m_i'+2)(l+l_i'+m+m_i'+3) < 0,$$

which is true iff $l + l_i' > m + m_i'$.

Step 2: The expected number of matching arguments between a speaker i, with speech characterized by σ, and a receiver j, characterized by $p(\theta|l_j', m_j')$, is increasing in σ if and only if $l_j' > m_j'$.

Note that $(l+m)$ is a binomially distributed random variable, where the number of trials is $(n - l_j' - m_j')$ and the probability that any one trial is a success is $(\theta_j\sigma + (1-\theta_j)(1-\sigma))$. Thus,

$$E[l+m|\sigma, \theta_j] = (n - l_j' - m_j')(\theta_j\sigma + (1-\theta_j)(1-\sigma)) \tag{14}$$
$$E[l+m|\sigma, p(\theta|l_j', m_j')] = \int_0^1 p(\theta|l_j', m_j')E[l+m|\theta, \sigma]d\theta. \tag{15}$$

Substituting (14) and (2) into (15) and differentiating with respect to σ yields

$$\frac{\partial E[l+m|\cdot]}{\partial \sigma} = (n - l_j' - m_j')[2E[\theta_j|p(\theta|l_j', m_j')] - 1],$$

which is positive iff $E[\theta_j|p(\theta|l_j', m_j')] > \frac{1}{2}$. Equivalently,

$$\begin{aligned} & E[\theta_j|p(\theta|l_j', m_j')] \\ = & \int_0^1 \frac{\theta^{l_j'+1}(1-\theta)^{m_j'}}{\int_0^1 \hat{\theta}^{l_j'}(1-\hat{\theta})^{m_j'}d\theta}d\theta = \frac{\Gamma(l_j'+2)\Gamma(l_j'+m_j'+2)}{\Gamma(l_j'+1)\Gamma(l_j'+m_j'+3)} = \frac{(l_j'+1)!(l_j'+m_j'+1)!}{l_j'!(l_j'+m_j'+2)!} \\ = & \frac{l_j'+1}{l_j'+m_j'+2} > \frac{1}{2}, \text{ or } \; l_j' > m_j'. \end{aligned}$$

Step 3:

From Step 1, $\text{Var}(\theta_j|p(\theta|l, k-l)) < \text{Var}(\theta_j|p(\theta|l-1, k-l+1))$ for all k and all $l \leq k$, iff $l + l_j' > m + m_j'$. From Step 2, $E[l+m|\sigma, l_j' > m_j']$ is increasing in σ. Higher σ implies higher $E[l]$ and lower $E[m]$. Thus, more weight is placed on events that correspond to posterior beliefs with lower variances as σ increases. From (7), it follows that the expected utility of the receiver with $l_j' > m_j'$ is increasing in σ in expectation. ■

Proposition 4

Proof Let $W = \{(l', m') : l' \geq 0, m' \geq 0, \text{ and } n > l' + m' > 0\}$. We show in Step 1 that for any $n \geq 3$, there exists a partition of W into W_1, W_2, W_3 s.t. (a) $\forall w \in W_1$, there exists $\hat{\sigma} > 0$ s.t. a listener whose active arguments are described by w prefers not to hear σ $\forall \sigma < \hat{\sigma}$; (b) $\forall w \in W_2$, a listener whose active arguments are described by w prefers to hear σ $\forall \sigma \in [0,1]$; and (c) $\forall w \in W_3$, there exists $\hat{\sigma} < 1$ s.t. a listener whose active

arguments are described by w prefers not to hear σ $\forall\sigma > \hat{\sigma}$. We then show in Step 2 that $W_1 \neq \emptyset, W_2 \neq \emptyset, W_3 \neq \emptyset$ for $n = 3$, and then extend the argument in Step 3 to any $n \geq 3$.

Step 1.

From Proposition 3, i's expected utility from hearing σ, $U_i(\sigma, l_i', m_i', n)$, is strictly monotonic in σ_j, thus crosses the expected utility from hearing nothing, $E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$, at most once on $\sigma \in [0,1]$. Suppose it does cross. From (2) - (7), $U_i(\sigma, l_i', m_i', n)$ is continuous in σ_j. Thus, by the intermediate value theorem, there exists a value of σ such that $U_i(\sigma, l_i', m_i', n) = E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$. Denote this value $\hat{\sigma}(l_j', m_j')$. From Proposition 3, $U_i(\sigma, l_i', m_i', n)$ is increasing in σ if and only if $l_j' > m_j'$; thus, if $l_j' > m_j'$, $U_i(\sigma, l_i', m_i', n) < E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$ $\forall\sigma < \hat{\sigma}(l_j', m_j')$ and if $m_j' > l_j'$, $\forall\sigma > \hat{\sigma}(l_j', m_j')$. Suppose now $U_i(\sigma, l_i', m_i', n)$ does not cross $E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$. Then, either $U_i(\sigma, l_i', m_i', n) > E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$ $\forall\sigma \in [0,1]$ or $U_i(\sigma, l_i', m_i', n) < E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$ $\forall\sigma \in [0,1]$. From (13), $E[u_j(\pi_j^*|p(\theta|l+l_j', m+m_j'))]$ is increasing in l for $l+l_j' > m+m_j'$ and in m for $l+l_j' < m+m_j'$. Thus, if $l_j' > m_j'$, $U_i(\sigma, l_i', m_i', n) > E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$ for $\sigma = 1$. It follows that $U_i(\sigma, l_i', m_i', n) > E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$ $\forall\sigma \in [0,1]$.

Step 2.

We next show that $W_1 \neq \emptyset, W_2 \neq \emptyset$, and $W_3 \neq \emptyset$ for $n = 3$.

$\forall\sigma \in [0,1]$, $\Pr((l,m) = (0,0)) > 0$ and $E[u_j(\pi_j^*|p(\theta|l_j'+0, m_j'+0))] = E[u_j(\pi_j^*|p(\theta|l_j', m_j'))]$. Consider $(l_j', m_j') = (2,0)$. From (12), $E[u_j(\pi_j^*|p(\theta|2,0))] = -\frac{3}{80}$. If $\sigma = 0$, then $(l,m) \in \{(0,0),(0,1)\}$. ($\Pr(l,m) = (1,0)|\sigma = 0) = 0$.) From (12), $E[u_j(\pi_j^*|p(\theta|2+0, 0+1))] = -\frac{1}{25}$. Since $-\frac{1}{25} < -\frac{3}{80}$, it follows that $(2,0) \in W_1$. By symmetry, $(0,2) \in W_3$.

Consider now $(l_j', m_j') = (1,1)$. Proceeding as above, $E[u_j(\pi_j^*|p(\theta|1,1))] = -\frac{1}{20}$. $(l,m) \in \{(0,0),(0,1),(1,0)\}$ $\forall\sigma \in [0,1]$. $E[u_j(\pi_j^*|p(\theta|1+0, 1+1))] = E[u_j(\pi_j^*|p(\theta|1+1, 1+0))] = -\frac{1}{25} > -\frac{1}{20}$. Hence, $(1,1) \in W_2$.

Step 3.

Let $n = 3+k$, $k \geq 0$. Let $(l_j', m_j') = (2+k, 0)$ and $\sigma = 0$. Again, $(l,m) \in \{(0,0),(0,1)\}$ and $E[u_j(\pi_j^*|p(\theta|2+k+0, 0+1))] < E[u_j(\pi_j^*|p(\theta|2+k, 0))]$. Thus, $(2+k, 0) \in W_1$, and symmetrically, $(0, 2+k) \in W_3$. For k even, consider $(\frac{k}{2}+1, \frac{k}{2}+1)$. At most one argument can be learned: $(l,m) \in \{(0,0),(0,1),(1,0)\}$. Since $-\frac{(\frac{k}{2}+3)(\frac{k}{2}+2)}{(k+6)(k+5)^2} > -\frac{(\frac{k}{2}+2)}{(k+5)(k+4)^2}$, it follows that $(\frac{k}{2}+1, \frac{k}{2}+1) \in W_2$. For k odd, consider $(\frac{k-1}{2}+1, \frac{k-1}{2}+1)$. Now as many as two arguments can be learned, and $(l,m) \in \{(0,0),(0,1),(0,2),(1,0),(1,1),(2,0)\}$. Proceeding as above, we obtain the following:

$$
\begin{aligned}
E[u_j(\pi_j^*|p(\theta|l_j'+0, m_j'+0))] &= E[u_j(\pi_j^*|p(\theta|l_j', m_j'))] = -\frac{(\frac{k-1}{2}+2)^2}{(k+4)(k+3)^2} \\
E[u_j(\pi_j^*|p(\theta|l_j'+0, m_j'+1))] &= E[u_j(\pi_j^*|p(\theta|l_j'+1, m_j'+0))] = -\frac{(\frac{k-1}{2}+3)(\frac{k-1}{2}+2)}{(k+5)(k+4)^2} \\
E[u_j(\pi_j^*|p(\theta|l_j'+1, m_j'+1))] &= -\frac{(\frac{k-1}{2}+3)^2}{(k+6)(k+5)} \\
E[u_j(\pi_j^*|p(\theta|l_j'+0, m_j'+2))] &= E[u_j(\pi_j^*|p(\theta|l_j'+2, m_j'+0))] = -\frac{(\frac{k-1}{2}+4)(\frac{k-1}{2}+2)}{(k+6)(k+5)^2}.
\end{aligned}
$$

Given $\dot{k} > 0$,

$$\forall z \in \{-\frac{(\frac{k-1}{2}+3)(\frac{k-1}{2}+2)}{(k+5)(k+4)^2}, -\frac{(\frac{k-1}{2}+3)^2}{(k+6)(k+5)}, -\frac{(\frac{k-1}{2}+4)(\frac{k-1}{2}+2)}{(k+6)(k+5)^2}\},\ z > -\frac{(\frac{k-1}{2}+2)^2}{(k+4)(k+3)^2}.$$

It follows that $(\frac{k-1}{2}+1, \frac{k-1}{2}+1) \in W_2$. ■

Proposition 5

Proof (1) The rightward policy shift by listener i is

$$\frac{l + l'_i + 1}{l + l'_i + m + m'_i + 2} - \frac{l'_i + 1}{l'_i + m'_i + 2} = \frac{l(m'_i + 1) - m(l'_i + 1)}{(l'_i + m'_i + 2)(l + l'_i + m + m'_i + 2)}.$$

Because each dimension is an independent draw from the same distribution, we can, without the loss of generality, restrict our attention to the expected result of deliberation on a single latent dimension. The expected rightward shift is, then,

$$\begin{aligned} &\Pr(l = 1 | l + m \in \{0, 1\}) \frac{m'_i + 1}{(l' + m'_i + 2)(l'_i + m'_i + 3)} \\ &+ \Pr(m = 1 | l + m \in \{0, 1\}) \frac{-l'_i - 1}{(l'_i + m'_i + 2)(l'_i + m'_i + 3)} \\ &+ (1 - \Pr(l = 1 | l + m \in \{0, 1\}) - \Pr(m = 1 | l + m \in \{0, 1\}))(0). \end{aligned} \tag{16}$$

$\Pr(l = 1 | l + m \in \{0, 1\}) = \sigma E[\theta_i]$ and $\Pr(m = 1 | l + m \in \{0, 1\}) = (1 - \sigma)(1 - E[\theta_i])$. Let $l'_i + m'_i = X$. We can re-write (16) as

$$\frac{1}{(X+2)(X+3)} [\sigma E[\theta_i](X - l'_i + 1) - (1 - \sigma)(1 - E[\theta_i])(l'_i + 1)]. \tag{17}$$

Given that $E[\theta_i] = \frac{l'_i+1}{l'_i+m'_i+2} = \frac{l'_i+1}{X+2}$, we can further re-write (17) as

$$\frac{1}{(X+2)(X+3)} \left[\frac{2\sigma - 1}{X+2} (l'_i + 1)(X - l'_i + 1) \right] \equiv \Delta(l'_i, X, \sigma).$$

Differentiating, we obtain $\frac{\partial \Delta(l'_i, X, \sigma)}{\partial l'} = \frac{2\sigma - 1}{(X+2)^2(X+3)}(X - 2l'_i)$, which is greater than 0 iff $l'_i < \frac{X}{2}$, i.e., iff $l'_i < m'_i$.

Part (2) of the proposition follows by symmetry. ■

Proposition 6

Proof We show (1) the benefits of speaking are greater for more extreme $E[\theta_i]$ than for more moderate $E[\theta_i]$; and (2) the benefits of listening are greater for more moderate $E[\theta_i]$ than for the more extreme $E[\theta_i]$.

(1) Suppose $\sigma^*_j > \frac{1}{2}$. From Corollary 1, the audience moves rightward in expectation. Given the optimality of σ^*_j and $\sigma^*_j > \frac{1}{2}$, j benefits in expectation from the expected move-

ment. Because marginal returns to proximity of others' policy choices to i's ideal point are diminishing, $\forall i$ s.t. $E[\theta_i] > E[\theta_j]$, holding constant the degree of uncertainty, i obtains greater expected benefits than j from the other players' policy changes in response to σ_j^*. From Lemma 2, $\sigma_i^* > \sigma_j^*$, and thus i's expected benefits from speaking must be greater still. The argument is symmetric for $\sigma_j^* < \frac{1}{2}$.

(2) From Proposition 5, if $E[\theta_i] > E[\theta_h] > \frac{1}{2}$, then hearing $\sigma > \frac{1}{2}$ induces a smaller change in π_i than in π_h, and thus is less beneficial to i than to h. The argument is symmetric for $E[\theta_i] < E[\theta_h] < \frac{1}{2}$ and $\sigma < \frac{1}{2}$. ■

Proposition 7

Proof Consider, without loss of generality, $i, h \in \mathcal{S}$ s.t. $\frac{1}{2} < E[\theta_h] < E[\theta_i]$ and $j \in \mathcal{R}$ s.t. $l'_j > m'_j$.

(1) j most prefers hearing speech s s.t. $s^t = 1\ \forall t \in \mathcal{L}'_j$. Thus, if $\exists j \in \mathcal{R}$ who chooses to hear h, then either $s_h^t = s_i^t = 1\ \forall t \in \mathcal{L}'_j$ or $\exists t \in \mathcal{L}'_j$ s.t. $s_i^t \neq 1$ and $t \in \Sigma_h^0 \cup \Sigma_h^1$. If $s_i^t \neq 1$ for some $t \in \mathcal{L}'_j$, then $\exists k$ s.t. $g_k = i$ and $E[\theta_i] < E[\pi_k|\Sigma_i^1 \cup \{t\}, \Sigma_i^0 \backslash \{t\}, g_k = i]$.

(2) Suppose $\forall k \in \mathcal{R}$ who choose to hear i, $\mathcal{L}'_j \cap \mathcal{L}'_k = \emptyset$. Then i's speech s_i^t is unconstrained $\forall t \in \mathcal{L}'_j$. Because $E[\theta_i] > E[\theta_h]\ \forall h \in \mathcal{S}\backslash i$, i prefers to induce a greater expected shift rightward by j than does $h \in \mathcal{S}\backslash i$. Because j prefers hearing right-biased speech to hearing left-biased speech, i can attract j by offering speech that induces at least as great an expected rightward shift as the speech offered by $h \in \mathcal{S}\backslash i$. Thus, if $\mathcal{L}'_j \cap \mathcal{L}'_k = \emptyset\ \forall k$ listening to i, then j also listens to i.

(3) Suppose $|\Sigma_i| = |\Sigma_h| = n$. $\exists j$ s.t. $g_j^* = h$ and $E[\theta_j] > \frac{1}{2}$. By Proposition 3, $\sum_{t \in \mathcal{L}'_j} s_h^t \geq \sum_{t \in \mathcal{L}'_j} s_i^t$. Because $E[\theta_h] < E[\theta_i]$, i prefers that j hear more right-biased speech than that preferred by h, hence both i and h are better off if j hears h.

(4) A. Suppose $|\mathcal{L}'_j \cap (\Sigma_i^0 \cup \Sigma_i^1)| \geq |\mathcal{L}'_j \cap (\Sigma_h^0 \cup \Sigma_h^1)|$. Then $g_j^* = h$ iff $|\mathcal{L}'_j \cap \Sigma_h^1| \geq |\mathcal{L}'_j \cap \Sigma_i^1|$ and $|\mathcal{L}'_j \cap \Sigma_h^0| \leq |\mathcal{L}'_j \cap \Sigma_i^0|$ because from Proposition 3, j prefers more right-biased speech. It follows that $E[\pi_j|\Sigma, g_j = h] \geq E[\pi_j|\Sigma, g_j = i]$. Given $E[\theta_i] > E[\theta_h]$, i benefits more than h from j's moving rightward, and prefers a greater move rightward than is induced in expectation by Σ_h. Thus, if h did not speak, i would choose $\hat{\Sigma}_i$ s.t.

(a) $\forall t \in \mathcal{L}'_j \backslash \cup_{\{k|g_k=i\}} \mathcal{L}'_k$, $t \in \hat{\Sigma}_i^1$; and

(b) $|\mathcal{L}'_j \cap \Sigma_h^1| \geq |\mathcal{L}'_j \cap \hat{\Sigma}_i^1| \geq |\mathcal{L}'_j \cap \Sigma_i^1|$ and $|\mathcal{L}'_j \cap \Sigma_h^0| \leq |\mathcal{L}'_j \cap \hat{\Sigma}_i^0| \leq |\mathcal{L}'_j \cap \Sigma_i^0|$.

From the optimality of Σ_i, $\exists k$ s.t. $g_k^* = i$, $\mathcal{L}'_j \cap \mathcal{L}'_k \neq \emptyset$, and $E[\pi_k|\hat{\Sigma}_i, g_k = i] > \max\{E[\theta_i], E[\pi_k|\Sigma_i, g_k = i]\}$. Thus, i weakly benefits from h's speaking (and strictly benefits if $|\mathcal{L}'_j \cap \Sigma_h^1| > |\mathcal{L}'_j \cap \Sigma_i^1|$).

B. Suppose $|\mathcal{L}'_j \cap (\Sigma_i^0 \cup \Sigma_i^1)| < |\mathcal{L}'_j \cap (\Sigma_h^0 \cup \Sigma_h^1)|$ and $g_j^* = h$. Because $E[\theta_i] > E[\theta_h] > \frac{1}{2}$ and players use undominated actions, $E[\pi_j|\Sigma, g_j = h] \geq E[\pi_j|\Sigma, g_j = i]$ implies i weakly benefits from h's speaking. Suppose $E[\pi_j|\Sigma, g_j = h] < E[\pi_j|\Sigma, g_j = i]$. Let

$$\hat{\Sigma}_i \in \arg \max_{\Sigma_i \in \{\Sigma_i | g_j(\Sigma_i, \Sigma_{-i}, l'_j, m'_j) = i\}} E[u_i(.)].$$

Because, ceteris paribus, j prefers to hear more extreme speech (from Proposition 3) and because $E[\theta_i] > E[\theta_h]$, $\Sigma_i^1 \subset \hat{\Sigma}_i^1$. Thus, if $g_k(\Sigma, l'_k, m'_k) = i$ then $g_k((\hat{\Sigma}_i, \Sigma_{-i}), l'_k, m'_k) = i$.

Let $x = |\mathcal{L}'_j \cap \hat{\Sigma}^1_i| - |\mathcal{L}'_j \cap \Sigma^1_h|$. $E[\theta_i] > E[\theta_h]$, hence $x > 0$ and $|\mathcal{L}'_j \cap (\hat{\Sigma}^0_i \cup \hat{\Sigma}^1_i)| \leq |\mathcal{L}'_j \cap (\Sigma^0_h \cup \Sigma^1_h)|$. Let $\check{\Sigma}^0_i$ be s.t. $|\mathcal{L}'_j \cap (\hat{\Sigma}^1_i \cup \check{\Sigma}^0_i)| = |\mathcal{L}'_j \cap (\Sigma^0_h \cup \Sigma^1_h)|$. Then,

$$E[\pi_j|\hat{\Sigma}^1_i, \check{\Sigma}^0_i, g_j = i] < E[\pi_j|\hat{\Sigma}^1_i, \hat{\Sigma}^0_i, g_j = i]. \tag{18}$$

Let $\hat{\Sigma}^1_h = \Sigma^1_h$ and $\hat{\Sigma}^0_h$ s.t. $|\mathcal{L}'_j \cap \hat{\Sigma}^0_h| \geq |\mathcal{L}'_j \cap \Sigma^0_h| - x$. Then

$$E[\pi_j|\hat{\Sigma}^1_i, \check{\Sigma}^0_i, g_j = i] - E[\pi_j|\Sigma^1_h, \Sigma^0_h, g_j = h] > E[\pi_j|\hat{\Sigma}^1_i, \check{\Sigma}^0_i, g_j = i] - E[\pi_j|\hat{\Sigma}^1_h, \hat{\Sigma}^0_h, g_j = h].$$

Combining with (18), we get

$$E[\pi_j|\hat{\Sigma}^1_i, \hat{\Sigma}^0_i, g_j = i] - E[\pi_j|\Sigma^1_h, \Sigma^0_h, g_j = h] > E[\pi_j|\hat{\Sigma}^1_i, \check{\Sigma}^0_i, g_j = i] - E[\pi_j|\hat{\Sigma}^1_h, \hat{\Sigma}^0_h, g_j = h].$$

Note that the right-hand side of this inequality is the expected rightward movement of π_j induced by speech of $\sigma = 1$ on x latent dimensions. There exists k s.t. $g^*_k = i$ and $\mathcal{L}'_j \cap \mathcal{L}'_k \neq \emptyset$. From the optimality of Σ_i, $g_j(\Sigma, l'_j, m'_j) = h$, and $E[\theta_i] > E[\theta_h]$, it must be that $E[\pi_k|\hat{\Sigma}^1_i, \hat{\Sigma}^0_i, g_k = i] > E[\theta_i]$. The expected rightward movement of π_k induced by replacing Σ_i with $\hat{\Sigma}_i$ is $E[\pi_k|\hat{\Sigma}^1_i, \hat{\Sigma}^0_i, g_k = i] - E[\pi_k|\Sigma^1_i, \Sigma^0_i, g_k = i]$, which is equivalent to the expected change in π_k induced by speech $\sigma = 1$ on x latent dimension.

From Proposition 5,

$$E[\pi_j|\hat{\Sigma}^1_i, \hat{\Sigma}^0_i, g_j = i] - E[\pi_j|\Sigma^1_h, \Sigma^0_h, g_j = h] > E[\pi_k|\hat{\Sigma}^1_i, \hat{\Sigma}^0_i, g_k = i] - E[\pi_k|\Sigma^1_i, \Sigma^0_i, g_k = i].$$

Because $E[\pi_k]$ is closer than $E[\pi_j]$ to $E[\theta_i]$, marginal diminishing returns ensures that i benefits more from j's move toward her than k's move away. Thus, if there are as many listeners like j as there are listeners like k, i prefers $\hat{\Sigma}_i$. Because $\hat{\Sigma}_i$ induces greater expected rightward movement of j, the benefit to i of offering $\hat{\Sigma}_i$ instead of Σ_i is greater still. Thus, Σ_i and $g^*_j = h$ imply that there are more listeners like k (choosing i) than like j (choosing h). ■

Catherine Hafer, Department of Politics, New York University
Dimitri Landa, Department of Politics, New York University

References

[1] Aragones, Enriqueta, et al. "Rhetoric and Analogies," (Tel-Aviv University Mimeo, 2001).

[2] Austen-Smith, David and Tim Feddersen, "Deliberation and Voting Rules." (Northwestern University Mimeo, 2002).

[3] Austen-Smith, David and Tim Feddersen, "Deliberation, Unanimity, and Majority Rule" (Northwestern University Mimeo, 2004).

[4] Baron, David P., "Private Politics," *Journal of Economics & Management Strategy* XII (2003), 31-66.

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

[6] Burnstein, Eugene and Amiram Vinokur, "Persuasive Argumentation and Social Comparison as Determinants of Attitude Polarization," *Journal of Experimental Social Psychology* XIII (1977), 315-22.

[7] Calvert, Randall and James Johnson, "Rational Actors, Political Argument and Democratic Deliberation." (University of Rochester Mimeo, 1998).

[8] Crawford, Vincent and Joel Sobel, "Strategic Information Transmission," *Econometrica* L (1982), 1431-51.

[9] Dawes, Robyn M., "Behavioral Decision Making and Judgment," *The Handbook of Social Psychology*, D. T. Gilbert, et al. eds., 4th ed. (New York: McGraw Hill, 1998).

[10] Dickson, Eric, Catherine Hafer, and Dimitri Landa, "Cognition and Strategy: a Deliberation Experiment." (New York University Mimeo, 2005).

[11] Gerardi, Dino and Leeat Yariv, "Putting Your Mouth Where Your Mouth Is: An Analysis of Collective Choice With Communication." (Yale University Mimeo, 2002).

[12] Glaeser, Edward L., Giacomo A. M. Ponzetto, and Jesse M. Shapiro, "Strategic Extremism: Why Republicans and Democrats Divide on Religious Values," *Quarterly Journal of Economics* CXX (2005), 1283-1330.

[13] Glazer, Jacob and Ariel Rubinstein, "On the Pragmatics of Persuasion: a Game Theoretical Approach." (Tel-Aviv University Mimeo, 2005).

[14] Hafer, Catherine and Dimitri Landa, "Deliberation as Self-Discovery and Institutions for Political Speech," *Journal of Theoretical Politics*, forthcoming.

[15] Hafer, Catherine and Dimitri Landa, "Deliberation, Ideological Bias, and Group Choice." (New York University Working Paper, 2005).

[16] Lanzi, Thomas and Jerome Mathis, "Argumentation in Sender-Receiver Games." (Mimeo, 2004).

[17] Lindzey Gardner and Elliot Aronson, eds. *Handbook of Social Psychology*, 3rd edition. (New York: Random House, 1985).

[18] Lipman, Barton and Duane J. Seppi, "Robust Inference in Communication Games with Partial Provability," *Journal of Economic Theory* LXVI (1995), 370-405.

[19] Lord, C. G., L. Ross, and M. R. Lepper, "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology* XXXVII (1979), 2098-2109.

[20] Lupia, Arthur, "Deliberation Disconnected: What It Takes to Improve Civic Competence," *Law and Contemporary Problems* LXV (2002), 133-50.

[21] Meirowitz, Adam, "In Defense of Exclusionary Deliberation: Communication and Voting with Private Beliefs and Values." (Princeton University Mimeo, 2004).

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

[23] Moscovici, Serge and Marisa Zavalloni, "The Group as Polarizer of Attitudes," *Journal of Personality and Social Psychology* XII (1969), 125-35.

[24] Murphy, Kevin M. and Andrei Shleifer, "Persuasion in Politics," *American Economic Review Papers and Proceedings* XCIV (2004), 435-39.

[25] Rabin, Matthew, "Psychology and Economics," *Journal of Economic Literature* XXXVI (1998), 11-46.

[26] Rabin, Matthew and Joel L. Schrag, "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics* (1999), 37-82.

[27] Rawls, John. 1971. *A Theory of Justice.* Harvard University Press.

[28] Shapiro, Andrew L., *The Control Revolution.* (New York: Perseus Publishing, 1999).

[29] Sunstein, Cass R., "The Law of Group Polarization," *Journal of Political Philosophy* 10 (2002), 175-95.

[30] Sunstein, Cass R., *Republic.com.* (Princeton: Princeton University Press, 2001).

[31] Wason, P. C.,"Self-Contradictions," *Thinking: Readings in Cognitive Science*, P. N. Johnson-Laird and P. C. Wason, eds. (Cambridge: Cambridge University Press, 1977).

[32] Wason, P. C., "Reasoning About a Rule," *Quarterly Journal of Experimental Psychology* XX (1968), 273-81.

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