Cognition and Strategy: A Deliberation Experiment

Eric S. Dickson* Catherine Hafer† Dimitri Landa‡

Abstract

A theory of deliberation must provide a plausible account both of individuals' choices to speak or to listen and of how they reinterpret their own views in the aftermath of deliberation. We describe a game-theoretic laboratory experiment in which subjects with diverse interests speak or listen before voting over a common outcome. An important feature of our strategic setting is that introspective agents may, upon hearing an unpersuasive argument, update away from the speaker's preferred position. While subjects are responsive to strategic incentives, they also deviate from Bayesian predictions by "overspeaking" when speech is likelier to alienate than persuade. Subjects thus come closer to the deliberative democratic ideal of a free exchange of arguments than equilibrium predictions suggest. We interpret evidence from subjects' deliberative choices and policy votes in terms of a cognitive hierarchy among subjects, defined by differing abilities to grasp the strategic implications of different kinds of information.

*Direct Correspondence to: Eric S. Dickson, Assistant Professor, Department of Politics, New York University. Email: eric.dickson@nyu.edu

†Assistant Professor, Department of Politics, New York University. Email: catherine.hafer@nyu.edu

‡Assistant Professor, Department of Politics, New York University. Email: dimitri.landa@nyu.edu

1 Introduction

Recent trends of thought in democratic theory have emphasized the positive effects that deliberation may have, via one mechanism or another, on the outcomes and legitimacy of collective decision-making (Cohen 1997; Elster 1997; Habermas 1990; Manin 1987). In particular, decisions made after exposure to individuals' "best arguments," freely exchanged during debate, may be more acceptable from a normative standpoint (Cohen 1997). Of course, deliberation can offer the full measure of such benefits only if agents in fact *choose* to communicate their arguments in the first place, an outcome that is often taken for granted in the normative literature.

In contrast, strategic choice of communications is an emphasis of the growing game-theoretic literature on deliberation. A number of studies delineate the circumstances under which individuals do – or do not – have an incentive to truthfully reveal private information to others in advance of a collective decision (Austen-Smith and Feddersen 2006; Calvert 2006; Gerardi and Yariv 2007; Meirowitz 2007; Patty 2005). Other studies model deliberation more explicitly as an exchange of arguments; perhaps counterintuitively, these studies find that individuals may prefer to remain silent rather than sharing their arguments under certain institutional and informational conditions (Glazer and Rubinstein 2005; Hafer and Landa 2007).

These two literatures offer distinctive perspectives on, and even definitions of, deliberation. Normative scholars have tended to define deliberation in terms of a particular set of behaviors, while game theorists have been inclined to define deliberation in terms of an institution or environment within which individuals make choices about mutual communication – and which may shape the nature of the choices that are made. The juxtaposition of these complementary approaches poses several key questions. Do deliberating agents make strategic choices in communicating with others – sharing arguments when they believe these will persuade others to change their minds, but withholding them at other times? Or do deliberating agents simply share their arguments freely, heedless of potential effects on future collective choices? And, in learning from deliberation, do agents "think strategi-

cally" about the implications of others' communications, or do they simply accept those arguments they deem valid while neglecting those they do not? This paper describes a laboratory experiment designed to offer an empirical perspective on these foundational questions.

Our framework takes as its starting point the observation that individual views on a given policy issue may be influenced by a number of *considerations* that the individual finds to be relevant – and that a specific consideration may be deemed relevant by some individuals but not by others. As an example, such considerations that might influence individual views on abortion policy include: (A) the principle of individual autonomy within one's own private sphere; (B) the intuition that criminalizing abortion might lead some women to pursue dangerous back-alley procedures that could lead to injury or loss of life; (C) the intuition that abortion-on-demand might lead to more unwanted pregnancies; and (D) the principle of the sanctity of life as depicted in scripture. A secular left-winger who is concerned about women's health, but who sees nothing morally wrong with abortion, may find (A) and (B) – but not (C) and (D) – to be considerations relevant to formulating a view on abortion rights. Conversely, a religious right-winger whose overriding concern involves the protection of "innocent life" may find (C) and (D) – but not (A) and (B) – to be relevant. At the same time, a pragmatic centrist might find the human costs implied by (B) and (C) to be relevant, but not the abstract moral principles of (A) or (D).

While individuals' views on a policy issue may potentially be influenced by a range of considerations, it is frequently the case, at a given point in time, that individuals do not fully appreciate the relevance of all these considerations. In our framework, agents choose either to speak or to listen before a collective decision made by simple majority rule. Individuals who choose to listen may, during the course of deliberation, hear arguments that bring to mind the relevance of "latent" considerations that had not previously appropriately factored into their judgments; individuals who choose to speak communicate an argument that could potentially "activate" a previously latent consideration in others' minds.[1] Because individuals can base their policy judgments - and ultimately their votes - only on those considerations whose relevance they actively have in mind, and not on latent considerations, the decision to speak or to listen is informed by a tradeoff between fleshing out one's own position

and attempting to influence the positions of others.

From a strategic standpoint, a key feature of our environment is that an attempt to activate a particular consideration in this way may induce a listener's policy judgments to adjust towards - or away from - the speaker's own. In the abortion rights example, consider an agent who initially has only consideration (B) in mind as being relevant to her judgment. Such an agent may be initially unsure whether to support an abortion rights proposal representing a "leftist" perspective (i.e., a policy influenced by considerations (A) and (B)) or one representing a more "centrist" perspective (i.e., a policy influenced by considerations (B) and (C)). Suppose that, during deliberation, such an agent chooses to listen to arguments advanced by others, and that a leftist counterpart communicates an argument that evokes consideration (A). If (A) was originally a "latent consideration" for the listening agent, then the leftist's argument brings the relevance of this consideration to the agent's mind, likely increasing her relative support for the leftist policy that was partly inspired by this consideration.[2] If, on the other hand, (A) is *not* a "latent consideration" for the listening agent – that is, if she does not find the principle of individual autonomy within one's own private sphere to be relevant after hearing the relevant arguments, either because she ultimately decides that the individual private sphere cannot be defensibly construed to include choices that affect the life of another being or that this principle is at best subordinate to other considerations – the communication will have failed *directly* to affect the set of considerations the listener has in mind when formulating a policy judgment. Yet, the possibility remains that an introspective listener may nonetheless update her policy views rationally given the *indirect* information implicit in the *failed* attempt to persuade – in a way that, in this example, would be unfavorable to the leftist speaker. Intuitively, a listener who reflects on her discovered indifference to consideration (A) should find it relatively less likely *ex post* that the leftist proposal, influenced by considerations (A) and (B), is preferable to the centrist proposal, influenced by considerations (B) and (C) – but not (A).

A famous literary example illustrates the indirect nature of this second kind of inference. In Arthur Conan Doyle's (1976) "Silver Blaze," Sherlock Holmes infers that a crime was not a burglary because the dog did not bark in the night. Having noted the dog's silence, Holmes' subsequent inference from

this fact was quite straightforward; the difference between Holmes and his companion Dr Watson is that only the great detective realized that the dog's failure to bark was itself an informative signal. Holmes and Watson differ in their abilities to make an inference from what might be called a null observation. In our deliberative framework, an unpersuasive argument – one that fails to engage a listener's latent consideration – similarly constitutes a null observation that can, nonetheless, be highly informative.

The extent to which individuals are, or are not, able to make such inferences in *interactive settings* is empirically a largely open question. A considerable literature in experimental psychology suggests that, in contexts of *individual decision making*, people tend to look for "positive" confirmations of hypothesized patterns, while disregarding or failing to look for "negative" information that does not fit the expected pattern but that could disconfirm the hypothesis (Wason 1968, 1977; Baron 1994 (Ch. 13); Taber and Lodge 2006). Such failures to use information optimally are often associated with cognitive biases by which individuals may update away from their previously held convictions only in light of direct and largely unambiguous evidence (Lord et al 1979; Zaller 1992;Rabin 1998; Baron 1994; Taber and Lodge 2006). Yet, the strategic, interactive environment of political deliberation is very different from the context of these studies of individual decision making. On the one hand, interactive environments are more complicated than non-interactive ones, suggesting a greater possibility for cognitive overload. On the other hand, strategic settings give individuals incentives to think through not only their own actions, but their counterparts' likely responses to those actions as well – and in an experimental setting such as ours, in which subjects gain experience in a variety of deliberative roles and receive extensive feedback about others' choices over a substantial number of periods, opportunities for learning and reflection on one's own ways of thinking may be further enhanced. As a result, the literature offers little reliable guidance as to the likely prominence of Bayesian and "Watsonian" ways of thinking in the psychology of deliberation.

The answer to this question has implications not only for the post-deliberation fit between agents' interests and their voting behavior, but for their chosen modes of deliberative engagement as speakers or listeners as well. Individuals who, as listeners, fail to learn from unpersuasive arguments will likely,

as speakers, fail to anticipate this insight in others – and accordingly will tend to speak even when this may be more likely to alienate than to persuade their listeners. In contrast, Bayesian agents would be expected to choose to share an argument only if, in expectation, such communication would do their cause more good than harm. As such, our experimental framework allows us to explore not only the microfoundations of judgment formation in deliberation, but also features of the aggregate behavioral patterns to which these microfoundations give rise.

Within the context of our experimental scenario, our results suggest that, at the aggregate level, deliberative behavior shows systematic deviations from the predictions of the Bayesian-agent benchmark. Agents who enter deliberation with a "moderate" active consideration – who, in our framework, have the most *ex ante* uncertainty about which policy alternative they prefer – overwhelmingly choose to listen, while agents with an "extreme" active consideration more often than not choose to speak. This difference in behavior across randomly-assigned experimental roles is in accordance with theoretical predictions, and suggests that strategic incentives do influence deliberative choices to a considerable extent. Yet, most subjects who begin with an "extreme" active consideration exhibit a strong and stable tendency to *overspeak* – that is, to speak much of the time even under circumstances in which they are more likely to alienate the favorably predisposed than to move listeners in their direction. Such endemic overspeaking exposes speakers to the risk of causing group outcomes to shift away from their private judgments – while in an overall sense enhancing the informational value of deliberation and bringing group behavior closer to the deliberative democratic ideal.

In order to account for these and other findings, we then turn to individual-level analyses linking a given subject's deliberative behavior with her ultimate *voting* behavior. We compare such individual-subject behavioral profiles not only to Bayesian-agent predictions but also to predictions we derive about two novel cognitive-behavioral ideal types: (1) *Overspeaking Watsonians*, who, like Watson but unlike the Bayesian Holmes, update their beliefs only when a latent consideration is explicitly activated; and (2) *Overspeaking Bayesians*, who are able to update their beliefs based on unpersuasive arguments, but who overspeak regardless, possibly because they fail to appreciate the strategic implications of others' ability to do the same. While a few subjects are found to behave in

a classically Bayesian manner, many more subjects behave in a manner consistent with one of these two types.

Our research contributes to a growing experimental literature on deliberation in particular, and political persuasion and communications more generally. Many existing studies explore behavior in highly contextual scenarios involving natural-language, "real issue" deliberation (e.g., Fishkin and Luskin 1996; Fishkin, Luskin, and Jowell 2000; French and Laver 2005). In contrast to these studies, our experimental scenario employs a purposely stylized setting; the impulse behind our design strategy is similar to that of Lupia and McCubbins (1998).[3] Both methodological approaches have distinct strengths and weaknesses that make them more or less suitable for exploring specific research questions related to deliberation. Natural-language studies are arguably indispensable for research into a variety of topics, such as detailed studies of attitudes on a specific political issue (e.g., Barabas 2004), or inquiries into the substance and style of arguments that are advanced in open discussion. Because we are interested in the extent to which individuals deviate from Bayesian rationality in deliberation and its aftermath, it is essential to employ an experimental design in which a clear benchmark for purely rational and outcome-oriented behavior can be defined. Only in the presence of such a benchmark can deviations from classical rationality be accurately identified. At a minimum, such a design requires a clear delineation of outcome-based interests and careful controls for the information to which individuals have access. A stylized, game-theoretic experiment is well-suited to such requirements; in the context of a "real issue" deliberation study, it would be extremely difficult if not impossible to measure with adequate precision subjects' prior attitudes, the range of arguments they *would* find persuasive *ex ante* (without activating these considerations in the course of measurement), subjects' beliefs about what might persuade others, and so on. In addition, by randomly assigning a given individual to different strategic positions over the course of a number of experimental periods, we are also able to make within-subject comparisons about how deliberative behavior changes across different circumstances.

2 The Model of Deliberation

The basic sequence of events in the model, reflecting the motivation in the Introduction, is as follows.[4] Individuals who begin in possession of *partial* information about their own best interests are given the opportunity to communicate with one another (the "deliberation stage"); once communication is complete, a vote is held between two potential outcomes, one of which is selected via simple majority rule (the "voting stage"). Each individual receives payoffs that depend on the degree of agreement between her *actual* individual best interests and the election-winning alternative.

Specifically, we consider deliberation within the context of a three-member group. Each member $i \in \{1,2,3\}$ has a type $t_i = (t_i^1, t_i^2) \in \{(A,B),(B,C),(C,D)\}$, where $A,B,C,D \in \mathbb{N}$ and $1 \leq A < B < C < D \leq 9$, and a "true number," $x_i^* = 10t_i^1 + t_i^2$, corresponding to her most-preferred outcome (e.g., ideal point). We use the notation $\overline{XY}$ to denote $10X + Y$ - that is, a two-digit number the first digit of which is X, and the second digit of which is Y. The ultimate social outcome is determined by majority rule over a pair of distinct alternatives, $\{y_1, y_2\}$, where $y_1, y_2 \in \{\overline{AB}, \overline{BC}, \overline{CD}\}$ are common knowledge among all group members from the beginning. An individual i's utility from outcome x is linearly decreasing in the "distance" between her true number and the outcome, $u_i(x, x_i^*) = c - |x_i^* - x|$, where c is a constant.

Deliberation has the potential to be persuasive because players do not know their true numbers for certain at the beginning of the game. Instead, each player initially possesses several pieces of information that are relevant to, but which do not necessarily uniquely determine, her true number. First, each player knows that the true numbers of every member of her group (including her own) must be drawn from a commonly known set $\{\overline{AB}, \overline{BC}, \overline{CD}\}$; it may be that all players have different true numbers from within the set, or it may be that two or more players share the same true number. Second, players know the (unconditional) probabilities corresponding to each of these true numbers - how likely it is that a given person would have each of the true numbers if they knew nothing else - all of which are positive. And third, each player knows a "fragment" of her true number, that is, one of the two digits of her true number - for example, "B" or "C" if her true number is $\overline{BC}$ - without

knowing whether that fragment is the first or the second digit of her true number. We refer to i's initially known fragment as her "active fragment" (or "active consideration"), $a \in \{t_i^1, t_i^2\}$, and to her initially unknown fragment as her "latent fragment" (or "latent consideration"), $l = \{t_i^1, t_i^2\} \backslash \{a\}$. The set of active fragments known to all group members is common knowledge.

Individuals' first strategic choice is their mode of deliberative participation $\lambda \in \{0, 1\}$, with $\lambda = 0$ capturing the decision to speak and $\lambda = 1$ the decision to listen.[5] A decision to speak entails an attempt to speak to all other members of her group; a decision to listen entails an attempt to listen to those other members of her group who have chosen to speak. Communication is successful only between individuals who have complementary modes of deliberation – e.g., if i speaks and j listens, communication from i to j takes place. Thus, anyone who chooses to speak receives no messages at all, and anyone who chooses to listen sends no messages at all. When complementary speaking and listening choices occur, j (the listener) receives a "message" m_j whose nature depends on both (1) speaker i's active fragment and (2) listener j's true number. If the speaker's active fragment is part of the listener's true number – corresponding either to the listener's active fragment or to her latent fragment – then the "message" listener j receives is simply speaker i's active fragment.[6] However, if the speaker's active fragment is not part of the listener's true number, then the "message" the listener receives is that she has received a "foreign fragment" – that is, she has received a fragment that is not a part of her true number (but that does not explicitly indicate what that fragment is).[7] Thus, given $\lambda_i = 0$ and $\lambda_j = 1$, $m_j = a_i$ if and only if $a_i \in \{t_j^1, t_j^2\}$; otherwise $m_j =$ "foreign."[8]

In our account of deliberation, receiving one's latent fragment corresponds to the activation of a consideration that had not previously played a role in one's policy judgment. Receiving a foreign fragment corresponds to the attempted priming of a consideration that the recipient does not consider relevant or valid. As foreshadowed in the Introduction, and detailed in the next section, either kind of communication can potentially be informative for its recipient.

3 Theoretical Predictions

Given the deliberative environment described in the previous section, how can individuals be expected to behave? Because we are interested in the microfoundations of judgment formation in deliberation, it is useful as a baseline to begin by analyzing the behavior of standard Bayesian agents. Before presenting formal behavioral predictions, we illustrate the nature of inference expected from Bayesian agents through the following examples.[9]

Example 1. Suppose it is commonly known that the set of possible true numbers in a three-player group is $\{13, 37, 79\}$ and that the group members' active fragments are 1, 3, and 3. It follows that a given player with the active fragment 3 may have either 13 or 37 as her true number, because both of these contain the fragment 3 (but the true number 79 does not, and so the group must not include a member with that true number). Suppose further that she chooses to listen, and is told that she has received a "foreign fragment." How can such a message prove informative for her? The other player with fragment 3 could not have sent the foreign fragment – if that player had sent her fragment, our listener would have been told that she had received the fragment "3" because "3" is part of her true number. Thus, the "foreign fragment" must have been sent by the subject with the fragment 1; hence the foreign fragment must have been 1; hence "1" must not be part of the listener's true number; hence her true number must be 37. ■

It is clear from this example that, in some circumstances, the receipt of a "foreign fragment" can be as informative as the receipt of the latent fragment. Further, the extent of deduction that is required in order to see this is not particularly demanding. The following two examples illustrate how the possibility of such inferences from the receipt of a "foreign fragment" affects optimal choices regarding speaking vs. listening.

Example 1a. As in the example described above, suppose that it is commonly known that the set of possible true numbers in the group of three players is $\{13, 37, 79\}$ and that the group members' active fragments are 1, 3, and 3, but also that the ultimate vote for a social outcome will be between 13 and 37. Suppose further that there is common knowledge within the group that 37 is more likely than

13; as such, in the absence of further information, the expected utility-maximizing choice for players with the fragment 3 is to vote 37 over 13. Consider first the incentives of the player with fragment 1 in choosing a deliberative strategy. If the players with fragment 3 both speak, the deliberative choice of the fragment 1 player is of no consequence. But that player is weakly better off speaking if at least one of the players with fragment 3 chooses to listen: if the latter receives a message indicating that 1 is her latent fragment, she will know that 13 is her true number, and will now vote for 13 instead of 37 – an improvement for the player with fragment 1 as this now ensures a majority for 13 over 37; if, instead, a listening fragment 3 player receives a message that she has received a foreign fragment, the player with fragment 1 is neither helped nor harmed. That is because a player with fragment 3 will learn that 37, not 13, is her true number *but she would have voted for* 37 *anyway* in the absence of communication because 37 is more likely than 13. Thus, it is a weakly dominant strategy for a player with fragment 1 to speak in this setting. ■

Example 1b. Suppose now exactly the same setting with one exception: it is commonly known within the group that 13 is a priori more likely than 37. The incentives facing the agents are now different. Sending fragment "1" cannot help, and it may hurt, since a Bayesian who receives a foreign fragment will choose 37. If the sender understands that the recipient may make such a deduction, she will strictly prefer not to send in such a situation. ■

We consider two different configurations of active fragments, in which players will face the incentives illustrated in the above examples. In configuration ABB, one player has active fragment "A" while two players have active fragment "B"; in configuration ABC, one player each has active fragment "A," "B," and "C". We denote the conditional probability that an agent with active fragment B has true number $\overline{AB}$ as $Pr(\overline{AB}|B)$. In either configuration, players will ultimately vote on $\overline{AB}$ vs. $\overline{BC}$. Players with active fragments A or C can be certain *ex ante* that their most-preferred point in the interval $[\overline{AB}, \overline{BC}]$ will be one of the endpoints ($\overline{AB}$ when A; $\overline{BC}$ when C), whereas players with active fragment B will have a true number whose *expected* value lies in the interior of that interval. In this context, we refer to the A and C active fragments as "extreme" and to the B active fragment as "moderate" (sometimes referring to the agents holding these active fragments as "extremists"

and "moderates," respectively). While many factors may distinguish "extremists" or "moderates" in different settings, our usage here captures two of the most commonly-cited: position on a left-right spectrum (at the endpoints vs in the middle), and a greater subjective sense of certainty on the part of extremists. Our conclusions about "extremist" and "moderate" behavior should be understood as being in relation to these two factors.

Our solution concept is weak dominance, which here generates unique equilibrium predictions with respect to both deliberation and voting. Table 1a summarizes the predictions about speaking and listening for players in different deliberative settings, while Table 1b summarizes the corresponding predictions for voting behavior, as a function of the messages received during deliberation.[10]

The logic of the examples induces broad patterns in the equilibrium predictions. When, as in Example 1a, Bayesian agents would perceive an incentive to speak, we will refer to this as the *Speaking Case* for the relevant A or C agent; when, as in Example 1b, such agents would perceive an incentive to listen (so as not to alienate favorably disposed counterparts), we will refer to this as the *Listening Case* for the relevant A or C agent.

4 Experimental Sessions

The experiment was carried out at the NYU Center for Experimental Social Science (CESS). Our results come from data collected in two experimental sessions involving 18 subjects each, for a total of 36 subjects. Subjects signed up for the experiment via a web-based recruitment system that draws from a broad pool of potential participants; individuals in the subject pool are mostly undergraduates from around the university, though a smaller number came from the broader community. We did not recruit from our classes, and all subjects gave informed consent according to standard human subjects protocols. Subjects interacted anonymously via networked computers; the experiment was programmed and conducted with the software z-Tree (Fischbacher 1999).

Each of our experimental sessions consisted of 30 rounds, where each round corresponds to a single play of our deliberation-and-voting game. Sessions lasted approximately 90 minutes, and on average

subjects earned US$26.56, including a showup fee of US$7. At the beginning of each session, a hard copy of the experimental instructions (which can be found in the reviewers' supplemental appendix) was distributed to each of the subjects; the instructions were also read aloud in an attempt to induce common knowledge. Subjects then took a six-question on-screen quiz to test their understanding and as a further means of inducing common knowledge of the deliberation framework; subjects were given immediate feedback as to the correct answers. After all 30 rounds were complete, subjects also completed a post-experiment debriefing survey.

At the beginning of each round, subjects were randomly rematched into a new group of 3 members and were given new fragment in a new deliberative environment. The matching of subjects and fragments was done in a quasi-random fashion designed to ensure wide experience of different strategic incentives. Because of our interest in the nexus of cognition and strategy in deliberation, and our desire to ensure that our inferences about subject behavior and cognition are not merely an artifact of inexperience in our deliberative setting, we devised an *ex ante* division of the session into two parts for the purposes of analysis. Subjects were to be deemed "inexperienced" during periods 1-12; the quasi-random assignment of subjects to fragment types and deliberative situations ensured that during this initial phase each subject was exposed three times each to Speaking Case (ABB), Listening Case (ABB), Speaking Case (ABC), and Listening Case (ABC), and had every type of fragment within each of these situations. This was done in order to ensure that, for each subject, many periods took place following this diverse exposure to every permutation of types of fragments and incentives, so that subjects' behavior when "experienced" could be measured. In accordance with our initial plan, most of our analyses thus concern the "experienced" periods 13-30, to strengthen our confidence that any inferences about behavior or cognition are not merely an artifact of the novelty of the scenario during the initial 12 periods of learning. This division of the session into two halves is for analysis purposes only; subjects were presented with the same interface, performed the same tasks, and received payoffs according to the same formula in all of the periods. Overall, each subject participated in the following distribution of situations: 7 Speaking Case (ABB); 7 Listening Case (ABB); 8 Speaking Case (ABC); 8 Listening Case (ABC). The specific parameters of each situation

(the set of possible true numbers, the unconditional probabilities, etc.) were different from round to round, but all groups in a given round were set in the same deliberative environment.[11] Subjects' payoffs were in cents, with $c = 80$, so that subjects received 80 cents in each round *minus* 1 cent for each "unit of distance" between the winning number and their own true number.

As feedback at the end of each round, subjects were told which alternative had won the vote (and the vote totals), their *actual* true number, and their payoff in cents. Such feedback allowed subjects to learn whether they had voted "correctly" (telling them something about how well they were using information), as well as offering some information about how their counterparts voted in the aftermath of communication.

5 Experimental Results: Aggregate Level

5.1 Deliberative Choices: Speaking and Listening

As the discussion in the theoretical results section demonstrates, individuals in our deliberation situation face differing strategic incentives depending on the active fragment that is known to them; the distribution of active fragments across other group members; and the relative likelihoods of different true numbers. Table 2a contains relevant data from our experimental sessions compiled in the same format as Table 1a.

The first thing to note is a systematic difference in deliberative behavior between agents depending on the nature of their active fragment – in particular, whether or not an individual's active fragment is more "moderate" or more "extreme." In the settings we describe, listening is always a weakly dominant strategy for individuals with moderate active fragments, while either speaking or listening can be weakly dominant for individuals with extreme active fragments, depending on the situation and their cognitive approach to deliberation.[12] Our first conclusion indicates that the distinction between the deliberative behavior theoretically expected of moderate and extreme agents is strongly apparent in our data.

Conclusion 1. *Subjects with more extreme active fragments speak more frequently than they*

listen, while subjects with more moderate active fragments listen more frequently than they speak.

The difference in deliberative behavior between "moderates" and "extremists" is striking, as can be seen in Table 2a. Over the last 18 periods of the experiment, once subjects had had some experience of each type of fragments and deliberative setting, subjects with a moderate ('B') active fragment chose their weakly dominant strategy – to listen – more than 96% of the time in each of the deliberative situations. Pooling across deliberative situations, they did so fully 98.1% of the time (306/312).[13] In contrast, subjects with an extreme ('A' or 'C') active fragment, whose incentives differed across different cases, chose to *speak* between 54.2% and 79.2% of the time in different deliberative situations over the last 18 periods of the experiment. Apart from being substantively interesting, this strong difference in the tendency of moderates and extremists to adopt different deliberative roles also stringly indicates that subjects' deliberative behavior was responsive to the details of their situations.

Our second conclusion reports the way in which their behavior varies between the Speaking Case and the Listening Case – that is, between situations in which a Bayesian agent would perceive that speaking as opposed to listening is her weakly dominant strategy.

Conclusion 2. *At the aggregate level, subjects with an extreme fragment chose to speak more often in the Speaking Case than they did in the Listening Case, with the speaking incidence on average substantially exceeding the predicted incidence for Bayesian agents. Learning over the course of the experimental sessions significantly increases the incidence of speaking in the Speaking Case but does not significantly alter the incidence of speaking in the Listening Case.*

Table 2a indicates that, in the last 18 periods, our subjects chose to speak between 54.2% and 66.7% of the time in each of the three settings corresponding to the Listening Case, for an overall average of 60.1% (101/168), in sharp contrast to the Bayesian prediction of 0%. Subjects spoke more often, between 73.3% and 79.2% of the time, in the three other settings corresponding to the Speaking Case, for an overall average of 76.8% (129/168), as against the predicted 100% for Bayesian agents. The null hypothesis that there is no difference in behavior between the Speaking Case and the Listening Case in these last 18 periods can be rejected decisively ($Z = 3.287$, $p < 0.001$). Thus, while predicted deliberative behavior for moderates conforms closely to the theoretical expectations

for Bayesians (as seen above, 98.1% observed versus 100% predicted listening), systematic deviations from the Bayesian predictions are observed for extremists, and by far the strongest deviations take the form of "overspeaking" by extremists in the Listening Case.

The importance and robustness of this "overspeaking" in the Listening Case is underscored by considering time trends in subjects' deliberative behavior. As shown in Table 2b, during the first 12 periods of the experiment, subjects in the Speaking Case chose to speak 62.0% of the time (67/108), whereas in the Listening Case they did so 63.0% of the time (68/108) – that is, there is no significant difference in deliberative behavior across Cases when subjects are inexperienced, but as shown above there is a substantial and significant difference once they are experienced. Further, it is important to note that the difference between the first 12 and the last 18 periods is due almost entirely to a shift in subjects' behavior in the Speaking Case. In the Speaking Case, the null hypothesis that behavior does not change from the first 12 periods (62.0%) to the last 18 periods (76.8%) is rejected ($Z = 2.636$, $p < 0.01$), whereas in the Listening Case, the null that behavior does not change from the first 12 periods (63.0%) to the last 18 (60.1%) cannot be rejected ($Z = 0.473$, $p = 0.64$). This observation strengthens confidence that subjects' failure to behave according to Bayesian predictions – and in particular, to exhibit a pronounced pattern of "overspeaking" in the Listening Case – is not entirely a result of inexperienced misunderstanding, because the behavior persists over the course of an experiment with a large number of rounds.

As noted in our theoretical predictions, we employ deliberative settings with different distributions of active fragments. This element of variation in our design gives us further ability to probe the extent to which subjects' *strategic* incentives influence their behavior, as opposed to other factors that are strategically irrelevant in our framework:

Conclusion 3. *Subjects with identical strategic incentives deliberate similarly across informationally distinct deliberative contexts.*

The ABB and ABC situations differ in descriptive and inferential complexity – a 'B' agent may have more difficulty discerning the meaning of a foreign fragment under ABC than under ABB, for example, while a 'C' agent with a strategic incentive to send may not see this so easily as a comparable

'A' agent, because the 'C' agent does not know her true number for certain while the 'A' agent does. Table 2a indicates that for experienced subjects in the last 18 periods, most of the variation in deliberative behavior across situations is captured by the strategic logic of the model rather than by such strategically irrelevant contextual factors. 'B' agents listen with remarkable consistency across deliberative settings – the variation between 96.7% and 99.0% is statistically insignificant. Similarly, 'A' and 'C' agents who, as Bayesian agents, would perceive an incentive to speak, also are remarkably consistent across contexts, doing so between 73.3% and 79.2% of the time, another insignificant difference. Finally, 'A' and 'C' agents who, as Bayesian agents, would perceive an incentive to listen, appear to do so a bit less consistently across contexts, varying from 54.2% to 66.7% of the time. In particular, such 'C' agents under ABC speak a bit less than do such 'A' agents under ABC, which may reflect 'C' agents' relative lack of true knowledge about their actual true number. However, even this difference is statistically insignificant.

Thus, subjects' strategic incentives matter more to their behavior than do such descriptive aspects of the deliberative problem as the degree of *ex ante* certainty they have about their true numbers. Even when listening is crucial to the subjects' ability to learn arguments necessary for determining their ideal policies (their unconstrained optima), their decisions regarding speaking and listening tend to be induced by the particular voting agendas - that is, by the extent to which what they could learn in deliberation is *necessary* for determining their "induced" preferences over the alternatives on the agenda (their constrained optima).

In order to further probe the extent to which behavior may have been affected by strategically superfluous descriptive features of the deliberative environment, we carried out a probit regression analysis of deliberative choice for those subjects with an extreme active fragment. Results from the regression, which was carried out for periods 13-30, are contained in Table 3. The dependent variable is the dichotomous choice between listening (1) and speaking (0). The independent variables depict a variety of features of the deliberative environment as well as a time trend variable. The regression results indicate that only subjects' strategic incentives – as categorized by the distinction between the Speaking and the Listening Cases – significantly affect behavior. Other factors, such as the

distinction between ABB and ABC situations; the distinction between "left-" and "right-handed" situations (e.g., ABB vs. CCD); and strategically irrelevant features of the true numbers' and probabilities' specific values all have a statistically insignificant effect on behavior. The time trend variable is also insignificant, indicating no strong learning trend after period 13. These results give us confidence that subjects' choices were affected by their strategic incentives but were not unduly influenced by other factors. With these results in mind, for the remainder of the section we separately pool all decisions together that are within a common case (Speaking or Listening), and restrict our attention to periods 13-30.

5.2 Voting Choices

We begin our consideration of the voting data by noting the frequencies with which subjects who have particular types of information (fragment) in a given round of the experiment are exposed to a variety of different deliberative outcomes. These frequencies are of course determined jointly by chance and by the deliberative strategies chosen by a given subject and his or her counterparts. For a given individual who has been assigned the 'B'-type fragment, three separate results of deliberation are about equally likely to be observed: the receipt of no fragment (29.9% of the time); the receipt of a foreign fragment only (26.7% of the time); and the receipt of the subject's latent argument only (25.0% of the time). The conjunction of the latent argument and a foreign fragment was considerably less frequent (15.7% of the time), and the rest of the possible outcomes essentially not present (from at or below 1% of the time).

The frequencies of the deliberative outcomes perceived by individuals who had been assigned the 'A'- or 'C'-type fragments reflect the different deliberative choices made by these agents compared to those who possess the 'B' fragment. Most of the time (84.2%) 'A' or 'C' agents receive no signal at all – naturally the case, given that 'B' agents almost always listen and the 'A' and 'C' agents themselves speak more often than they listen. 'A' and 'C' agents receive only a foreign fragment 13.6% of the time and observed other outcomes very rarely.

Along with our theoretical expectations, these distributions of deliberative outcomes lead us to

expect that we will learn the most about subjects' behavioral types from voting behavior by observing the choices made by the 'B' agents. Our next two conclusions focus on those choices. The first conclusion checks how well the subjects grasp the problem that is posed to them and on their ability to make the elementary probabilistic inferences they are expected to perform:

Conclusion 4. *Subjects almost always use dominant voting strategies when receiving no signal or when receiving their latent fragment.*

The data supporting this Conclusion can be found in Table 2c. Aggregating across deliberative settings, subjects with a 'B' (moderate) active fragment who receive no signal at all vote in accordance with their prior belief 94.9% of the time (150/158). Further, subjects with a 'B' active fragment who receive their latent fragment vote correctly (for their now-known true number) 95.0% of the time (209/220). Taken together, these statistics indicate a striking degree of understanding among our subjects both of the meaning of the unconditional probabilities they were given and of the meaning of latent fragments. This finding offers perspective relevant to our next conclusion, about voting behavior by those 'B' agents who receive only a foreign fragment:

Conclusion 5. *Subjects often fail to learn from informative foreign fragments.*

This conclusion exploits the differences in the information content of foreign fragments in different deliberative settings. In the Speaking Case (ABB), the 'B' agent's prior indicates that, in the absence of further information, she should prefer to vote for $\overline{BC}$ over $\overline{AB}$. If, in this setting, a Bayesian 'B' agent receives a foreign fragment, she will understand that the foreign fragment must have been sent by the 'A' agent; that her own true number cannot contain the 'A' fragment; and therefore that her own true number must be $\overline{BC}$. The same is also true in an ABC situation: a Bayesian agent who understands the strategic incentives faced by 'A' and 'C' types will expect that, if one foreign fragment only is received, that it would have most likely been sent *by the agent who was trying to change her mind* – and therefore should not affect her ultimate vote.

However, the situation in the Listening Case (ABB) is quite different. The 'B' agent's prior belief inclines her to vote for $\overline{AB}$ over $\overline{BC}$. However, unlike in the cases discussed above, receiving a foreign fragment puts the 'B' agent's prior beliefs in tension with the information represented by it. She now

ought to vote against her prior belief, because she would know with certainty that her true number could not contain fragment 'A.' As such, a comparison of the voting behavior of 'B' agents who have received only a foreign fragment between the Listening Case (ABB) and the other cases provides a direct test for the hypothesis that agents fail to make inferences from unpersuasive but informative arguments. The results of this comparison can also be found in Table 2c.

The rate with which subjects fail to vote "correctly" upon receiving a foreign fragment that indicates against their prior is striking. Taken over all 30 periods, subjects vote correctly against their prior 56.2% of the time (18/32) in the Listening Case (ABB), but they vote correctly with their prior 91.7% of the time (100/109) in the other cases. The null hypothesis that these success rates are identical is soundly rejected by a difference-of-proportions test ($Z = 4.778$, $p < 0.0001$). Restricting attention to the last 18 periods (13-30), subjects vote correctly (against their prior) 62.5% of the time (10/16) in the Listening Case (ABB), but they vote correctly (with their prior) 93.4% of the time (71/76) in the other cases. In this instance the null hypothesis that these success rates are identical is also overwhelmingly rejected ($Z = 3.465$, $p < 0.001$).

It is worth re-emphasizing that the high rate of errors in voting that we note above takes place in the context of the Listening Case (ABB). If a 'B' subject receives a foreign fragment it is *only* the 'A' agent from whom the foreign fragment could have come. While the information implicit in the receipt of such a foreign fragment is not as transparently *labeled* as the information implicit in the receipt of a latent fragment, it is in actuality no less informative.

6 Re-examining Microfoundations

The five conclusions presented above all concern aggregate-level analyses on data pooling together all of our experimental subjects. Of course, such aggregate-level results could result from a wide variety of different distributions of *individual*-level behavior; in this section, we report several analyses carried out at the individual level.

As we noted in the Introduction, we ultimately organize the data by supposing that different

individuals may understand the epistemic and strategic implications of foreign fragments – failed attempts at persuasion – to different degrees. We conceptualize cross-subject heterogeneity in this regard by defining several distinct cognitive-behavioral ideal types, and classifying subjects based on the degree to which their behavior is consistent with expectations about how members of such ideal types would behave. Before proceeding with this plan, we first consider, and reject, another account that might at first blush seem consistent with the evidence of overspeaking in the previous section. Suppose that a population of agents exists that has no or a limited understanding of the unconditional probabilities associated with the true numbers. As voters, such agents may have little idea how to cast their ballots when their initial active fragment is moderate and when their latent fragment is not activated during deliberation; perhaps they would simply vote randomly. However, these agents would be highly likely to vote correctly if their latent fragment *were* activated – with both fragments known, the probability information would become irrelevant. If such agents were present in sufficient numbers, a Bayesian with an extreme active fragment might have an incentive to speak – even in the Listening Case – if she believed that the likelihood of persuading such agents exceeded the risk of alienating Bayesian listeners.

However, this alternative account is implausible for several reasons. First, subjects overwhelmingly gave correct answers to our pre-experiment quiz question about the unconditional probabilities (34/36, 94.4%).[14] Second, as noted in Conclusion 4, subjects with a moderate active fragment who receive no signal at all practically always vote in a way consistent with the unconditional probability information, further suggesting that very few if any subjects possess the particular deficiency in understanding described above that might give Bayesian agents incentives always to speak. Finally, subjects' responses to our post-experiment survey provide further evidence that few subjects imagined others to understand the scenario more poorly than they themselves did.[15]

Given the implausibility of this alternative explanation, we proceed by defining two novel, analytically plausible cognitive-behavioral ideal types, deriving predictions as to how such agents could be expected to behave in our deliberative and voting stages, and comparing these predictions to the individual behavioral profiles of our experimental subjects. These novel types of agents stand in

contrast to, while sharing some features of, the standard Bayesian agents whose expected behavior was captured in our initial theoretical predictions. Such Bayesian agents, who correctly update their beliefs upon receiving either their latent fragment or notice of a foreign fragment, may be thought of as occupying the top step of a hierarchy of cognitive types, each of which possesses a different degree of understanding of the implications of information in the deliberative environment.[16]

Agents below the top step in such a hierarchy differ from Bayesian agents in their ability to perform one or more kinds of inference. Our earlier discussion motivates the definition of our first novel ideal type. Like Dr. Watson, these agents fail to learn from null observations – in our context, they are not capable of learning from the failure to persuade that results in a foreign fragment being delivered.[17] Because Watsonians do not grasp the significance of foreign fragments as voters, it seems natural to assume that they are unlikely to foresee as speakers that *other* voters may learn from foreign fragments. As such, we refer to this first ideal type as *Overspeaking Watsonian.* While such agents fail to learn from foreign fragments, they differ from Bayesians only in this respect; in all other ways, they update beliefs and understand the deliberative setting just as Bayesians do.

This depiction points to the possibility of a second novel ideal type which falls between Bayesians and Watsonians in a cognitive hierarchy. Our experimental game consists of a deliberative and a voting stage. While it seems natural to assume that agents who do not understand foreign fragments as voters are unlikely to foresee as speakers that *other* voters may glean information from them, it is more intuitive to imagine that there may exist other agents who *do* grasp the informational value of foreign fragments as voters, but who as speakers *do not* foresee that other voters may make the same inference. Such agents would be capable of making Bayesian inferences themselves as listeners, but would fail to appreciate the possibility that their speech might alienate others. As such, we refer to this type of agent as Overspeaking Bayesian. Overspeaking Bayesians also differ from Bayesians only in the specific way described.

Of course, other ideal types aside from Bayesians, Overspeaking Bayesians, and Overspeaking Watsonians could also be defined. Other kinds of agents could fall below Watsonians in a cognitive hierarchy; some agents may deviate more strongly from Bayesian inference, while others may simply

behave randomly. As we do not possess strong intuitions about other, specific types that might be present in our subject population, we provide no further type characterizations and simply refer collectively to other kinds of agents as *Deviant*.

As indicated above, neither Overspeaking Watsonian nor Overspeaking Bayesian agents conceive of the possibility that they may alienate fellow group members who, as a result of communication from the Overspeaking Watsonian/Overspeaking Bayesian, receive a foreign fragment. As a result, both Overspeaking Watsonian and Overspeaking Bayesian agents would be indifferent between sending and receiving whenever they have an extreme active fragment and moderate agents' prior beliefs are stacked in their favor. These are the circumstances which are represented in boldface in Table 1a. These ideal types' willingness to speak in such circumstances marks the only difference between their perceived optimal deliberative strategies and those of Bayesian agents.[18] Overspeaking Watsonians and Overspeaking Bayesians would behave differently only in the voting stage, never in the deliberation stage – and even in voting stage, only under Listening Case (ABB), upon receipt of a foreign fragment (the contingency given in boldface in Table 1b). In such a circumstance, Bayesians and Overspeaking Bayesians would vote for $\overline{BC}$ against their prior belief, while Watsonians would continue to prefer $\overline{AB}$.

Given these predictions, we can now look for evidence for each of these ideal types in the individual-level data. Keeping in mind the observational equivalence between Overspeaking Bayesians' and Overspeaking Watsonians' deliberative choices (i.e., speaking vs. listening decisions), we can state the following conclusion:

Conclusion 6. *Although there is substantial heterogeneity in the deliberative behaviors demonstrated by subjects when assigned to the position of extremist, the behavior theoretically associated with the Overspeaking Watsonian and the Overspeaking Bayesian behavioral types accounts for roughly half of the observations.*

Table 4 contains the individual-level deliberation choices in rounds 13-30 for subjects who possess extreme active fragments. (Note: subjects 1-18 took part in Session 1; subjects 19-36 took part in Session 2.) We restrict our attention to rounds 13-30 in an attempt to obtain a relatively meaningful,

if rough, classification of behavior that takes into account an initial period of learning by subjects.[19] In order to organize this data, we employ the following system of categorization. If a subject behaves in precisely the same way as a given behavioral type would, or deviates from this behavior by no more than one choice in the Speaking Case and by no more than one choice in the Listening Case, she is categorized as belonging to that behavioral type. For the Bayesian type, this involves speaking in all or all but one of the Speaking Case circumstances, and listening in all or all but one of the Listening Case circumstances. For the Watsonian and the Overspeaking Bayesian types, this involves speaking in all or all but one of the Speaking Case circumstances, and speaking in at least half of the Listening Case circumstances.[20]

This system of categorization yields the following results for our 36 subjects. 19 subjects can be uniquely categorized as belonging to the Overspeaking Watsonian/Overspeaking Bayesian types, while only 3 subjects can be uniquely categorized as belonging to the Bayesian type.[21] If the classification criterion had been slightly looser, allowing for one further deviation from expected ideal type behavior, 1 further subject would have been categorized as Bayesian, and 1 further subject would have been categorized as Overspeaking Watsonian/Overspeaking Bayesian; in Table 4, these subjects are labelled as "leaning" towards the relevant ideal type. 2 further subjects, who were simultaneously only one choice away from the criteria for Bayesian *and* for Overspeaking Watsonian/Overspeaking Bayesian classification, were left uncategorized. These classifications account for 26 of the 36 subjects; the remaining ten subjects, labelled as "deviant" in Table 4, did not fit a pattern consistent with any of our ideal types, sometimes engaging in especially perverse behavior – for example, always listening, regardless of the Case – or exhibiting a pattern of choices that appears essentially random.

In addition to these categorizations and data on deliberative behavior, Table 4 also contains data on subjects' voting choices under Listening Case (ABB) when only a foreign fragment was received. The next conclusion relates subjects' voting behavior with our above subject categorizations:

Conclusion 7. *At the individual level, subjects' deliberative behavior is highly correlated with their voting behavior upon receipt of only a foreign fragment: subjects who deliberate like Bayesian agents vote correctly; subjects who deliberate like Overspeaking Watsonian/Overspeaking Bayesian*

agents vote correctly less often; and subjects who deliberate like Deviant agents vote correctly least often.

If the intuition behind our cognitive-behavioral ideal types is correct, it should be the case that the reasoning processes we separately infer from first-stage deliberative choices and second-stage voting choices should be consistent within ideal type sub-populations. In order to test for such consistency, we related subjects' individual-level vote choices for the 32 instances of the Listening Case (ABB) contained in Table 2c with the ideal-type classifications we derived from these subjects' deliberative choices, made in different periods, that are depicted in Table 4. Agents classified as Bayesian based on deliberative behavior voted correctly 100% of the time (4/4); agents classified as Overspeaking Watsonian or Overspeaking Bayesian voted correctly 57.9% of the time (11/19); and agents classified as Deviant voted correctly 28.6% of the time (2/7). The pattern suggested by this result is consistent with the existence of stable cognitive-behavioral ideal types who are heterogenous in the degree to which they understand the informational content of foreign fragments. However, because the conjunction of events required to generate such voting observations occurs relatively infrequently, it is not possible to make strong statistical claims about the significance of this pattern.[22]

While subjects infrequently confront circumstances that allow individual classification based on voting behavior, the aggregate evidence that is available suggests that Overspeaking Watsonian and Overspeaking Bayesian types are likely both present in significant numbers. The subjects who are classified as Overspeaking Watsonian or Overspeaking Bayesian based on their deliberative behavior vote correctly in 57.9% (11/19) of such circumstances; if subjects were homogeneously Overspeaking Watsonian (vs. Overspeaking Bayesian) this figure would be 0% (vs. 100%). As we have seen, subjects nearly always vote correctly outside of Listening Case (ABB); the single, sharp downward deviation in voting accuracy in exactly the place it would be expected of Watsonians does seem to indicate their presence in the sample. However, while 57.9% is far from 100%, it is also far from 0%, pointing to the likely presence of Overspeaking Bayesians as well. Subject responses to our post-experimental survey questions provide further suggestive evidence along these lines.[23]

7 Conclusion

This paper has presented the results of a laboratory experiment exploring the nexus of cognition and strategy in deliberation. Within the context of a novel experimental scenario, in which a clear benchmark for fully Bayesian-rational behavior can be discerned, we find that subjects' deliberative behavior is responsive to strategically-relevant features of the environment. Subjects who are randomly assigned to more extreme initial positions tend to speak, while subjects who are randomly assigned to more moderate initial positions tend to listen, consistent with theoretical predictions. Yet, both at the aggregate and at the individual level, behavior deviates systematically from Bayesian predictions. The most striking result is a robust tendency towards overspeaking; subjects whose active consideration is extreme often speak even when, in expectation, such speech is more likely to alienate listeners than to persuade them. Our individual-level analysis takes advantage of the experimental design by studying subjects' deliberative behavior and vote choice across the different strategic and informational settings they encounter during different rounds of the experiment. The results suggest that some subjects – our Overspeaking Watsonians – likely overspeak because they do not understand the information implicit in failed attempts to persuade. Our Overspeaking Bayesians, on the other hand, do appear to understand the implications of such information – while overspeaking nonetheless. One plausible interpretation is a cognitive one as well – such agents may understand that inference from a null observation is possible, but may not realize that their own speech may generate null observations that can be interpreted by others.

Regardless of its source, overspeaking increases the amount of information publicly available for post-deliberative decision making – a social silver lining to behavior that appears to be, from the point of view of individual incentives, suboptimal. At the same time, our results also suggest that while some agents may take advantage of this "unexpected" informational windfall, others – in our experiment, almost half of the participants – may not.

The phenomenon of overspeaking may also be taken as empirical evidence bolstering the prospects of deliberation-based accounts of democracy. While our subjects do respond to strategic incentives,

they nonetheless come closer to the normative ideal of a free "exchange of arguments" than Bayesian-rational predictions would expect. It is particularly striking that this should be the case even in a highly stylized experimental environment, in which individuals' outcome-based interests can be quite clearly discerned. This outcome, while suggestive, can only achieve its fullest explication in the context of future work.

References

Austen-Smith, David and Tim Feddersen. 2006. "Deliberation, Preference Uncertainty, and Voting Rules." *American Political Science Review* 100(2), 209-217.

Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98(4), 687-701.

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

Binmore, Ken. 1990. *Essays on the Foundations of Game Theory.* London: Basil Blackwell.

Calvert, Randall L. 2006. "Deliberation as Coordination Through Cheap-Talk." Washington University Mimeo.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119(3), 861-898.

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

Costa-Gomes, Miguel, Vincent P. Crawford, and Bruno Broseta. 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica* 69(5), 1193-1235.

Doyle, Arthur Conan. 1976. *The Complete Sherlock Holmes.* Doubleday.

Elster, Jon. 1997. "The Market and the Forum: Three Varieties of Political Theory." In J. Bohman and W. Rehg, eds., *Deliberative Democracy.* Cambridge: MIT Press, 3-34.

Fischbacher, Urs. 1999. "z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual." Working Paper Nr. 21, Institute for Empirical Research in Economics, University

of Zurich.

Fishkin, James, and Robert C. Luskin. 1996. "The Deliberative Poll: A Reply to our Critics," *Public Perspective* 7(1), 45-49.

Fishkin, James, Robert C. Luskin, and Roger Jowell. 2000. "Deliberative Polling and Public Consultation," *Parliamentary Affairs* 53, 657-666.

French, Damien, and Michael Laver. 2005. "Participation Bias and Framing Effects in Citizens' Juries." Paper, Annual Meeting of the American Political Science Association.

Gerardi, Dino and Leeat Yariv. Forthcoming. "Deliberative Voting." *Journal of Economic Theory.*

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

Habermas, Jurgen. 1990. *Moral Consciousness and Communicative Action.* Cambridge: MIT Press.

Hafer, Catherine and Dimitri Landa. 2007. "Deliberation as Self-Discovery and Institutions for Political Speech." *Journal of Theoretical Politics*, forthcoming.

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

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

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

Lupia, Arthur and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need To Know?* Cambridge: Cambridge University Press.

McCubbins, Mathew D. and Daniel B. Rodriguez, 2006. "When Does Deliberating Improve Decision Making?" UCSD Mimeo.

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

Meirowitz, Adam. 2007. "In Defense of Exclusionary Deliberation: Communication and Voting with Private Beliefs and Values." *Journal of Theoretical Politics*, forthcoming.

Patty, John. 2005. "Arguments-Based Collective Choice." Harvard University Mimeo.

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

Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3), 755-769.

Wason, P. C. 1968. "Reasoning About a Rule." *Quarterly Journal of Experimental Psychology* 20, 273-81.

Wason, P. C. 1977. "Self-Contradictions." In P. N. Johnson-Laird and P. C. Wason, eds., *Thinking: Readings in Cognitive Science.* Cambridge University Press, 114-28.

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

Notes

[1]Communication in our framework thus stands in contrast to cheap talk models of deliberation, in which speakers send messages whose credibility (or lack thereof) is endogenous. Here, successful communication brings to mind the relevance of considerations that the recipient accepts as valid on their own "merits" (independent of the identity of the person who evokes them), but which she had not fully appreciated before that communication.

[2]For example, before deliberation, a person may be aware that consideration (D) in our example exists, without being cognizant of the relevance of that consideration for her position on the issue of abortion rights – as if saying "so what?" or "what is it to me?" In an alternative interpretation, a person may be aware that a religious argument against abortion exists without knowing what that religious argument actually says. These two interpretations are analytically equivalent in the context of our model, and both are consistent with our analysis. In the interests of clarity, we invoke only the former one throughout the paper.

[3]Though the substantive focus differs; Lupia and McCubbins study the conditions under which

an informationally-privileged agent can persuade, or deceive, her principal. In our model, speakers persuade by evoking considerations that listeners innately recognize as valid and relevant towards forming the judgment at hand. Our substantive results on deliberative choice are, to our knowledge, new to the literature. We also note that we have recently become aware of further experimental evidence for overspeaking in a deliberative setting in work by McCubbins and Rodriguez (2006).

[4]The supplemental appendix for reviewers contains the instructions given to subjects, demonstrating the way in which the experimental scenario was presented in the laboratory.

[5]Modeling λ as a binary choice between speaking and listening captures in a stark form a common feature of large- and small-scale deliberative processes, in which advocating a policy position by making supporting arguments tends to come at the cost of lessened contemplation of others' arguments. Among other reasons, such tradeoffs may exist because of cognitive limitations involving scarce attentional resources or imperfect memory, as emphasized in psychologically-oriented studies of the determinants of persuasion (Zaller 1992; Lupia 2002).

[6]For clarity, we note that by this assumption a speaker can only ever attempt to send a message corresponding to her active fragment – the consideration she actually has in mind pre-deliberation.

[7]Not showing to the receivers which fragment the received "foreign fragment" is allows us to approximate in the experiment a fundamental asymmetry between how agents learn from intrinsically persuasive arguments and how they may potentially learn from unpersuasive ones in "real world" deliberation. Whereas the meaning of an intrinsically persuasive argument is entailed in the argument itself, learning from unpersuasive arguments requires indirect strategic reasoning that makes inferences from, *inter alia*, what sort of argument the speaker in question would really like her audience to believe is persuasive or unpersuasive, etc. As we explain below, in the experiment, listeners are given all the information that is necessary, using analogous reasoning, to make the appropriate indirect inference that could be made in more contextually embodied deliberation.

[8]In the formal theory literature, our deliberative environment most closely resembles Hafer and Landa (2007), in which different individuals are also associated with distinct sets of considerations that can be "activated" in the course of debate – they, however, do not analyze subsequent collective

decision-making. Lipman and Seppi (1995) and Glazer and Rubinstein (2005) offer related models, though ones in which a given message is assumed to be equally persuasive to all.

[9]Later, after presenting some experimental results, we proceed to consider explicitly other, non-Bayesian cognitive-behavioral ideal types to provide an account of the microfoundations of deliberation that is more consistent with the observations in our experimental data.

[10]Formal arguments that these strategies are weakly dominant are contained in the reviewers' supplemental appendix that we will make available on-line in the event of publication.

[11]We also instantiated configurations CCD and BCD, but because these are strategically equivalent to ABB and ABC respectively, we subsume CCD into references to ABB and BCD into references to ABC except where specifically noted. Our choice of deliberative situations allows us to minimize potential interpretive confounds by allowing us to control for a variety of factors that are strategically irrelevant in our model but which could plausibly affect subject behavior – e.g., whether subjects know their true numbers for certain (when this is irrelevant to optimal strategies) or behave differently when their active fragments are on the left/lower vs on the right/higher. We do not employ disconnected configurations such as ABD because persuasion is possible only between agents with adjacent active fragments; at least one agent's deliberative choice is meaningless in such settings.

[12]Individual decisions are the unit of analysis throughout the paper unless otherwise noted.

[13]And four of the six instances in which a moderate spoke during these periods came from a single subject. Note that the corresponding figure over all 30 rounds is 94.9% (501/528).

[14]The question was: "Suppose that the set of possible true numbers is 26,68,89. Suppose that the frequency of the true number 26 is 35%, that the frequency of the true number 68 is 45%; and that the frequency of the true number 89 is 20%. And suppose that a person in your group is told that 6 is a fragment of his or her true number. What is his or her most likely true number given this information?" As feedback, subjects saw their own answer and the text: "The correct answer was 68. If the person has a fragment 6, his or her true number must be either 26 or 68 because 89 does not contain the fragment 6. And 68 occurs with greater frequency than 26."

[15]Note for now that, when asked "Did you find the problem at hand difficult or easy?" and "Do

you think other people found the problem difficult or easy?" only 4 of the 36 subjects (11.1%) gave responses indicating a belief that they found it easier than their counterparts, and only one of these 4 subjects exhibits deliberative and voting behavior that is consistent with the alternative explanation. As such, these responses weigh strongly against an interpretation of the data in which Bayesian subjects always attempt to activate the latent fragments of other subjects whose capacities are believed to be much lower. See also footnote 23 below.

[16]Some recent economics studies usefully interpret experimental findings in the context of such a cognitive hierarchy (Camerer, Ho, and Chong (2004); Costa-Gomes, Crawford, and Broseta (2001).

[17]In technical parlance, Watsonians are not logically omniscient because they violate *negative introspection* in updating their beliefs (see e.g. Binmore 1990, pp. 108-110).

[18]A formal derivation of Overspeaking Watsonians' optimal strategies is contained in the reviewers' supplemental appendix. Overspeaking Bayesians' optimal strategies are the same and are justified by the same arguments.

[19]The nature of the following discussion, and even the specific classifications, would be practically unchanged using data from all periods. As noted, the Table 3 regression contains no evidence of aggregate learning in Periods 13-30. Additional regression specifications attempted to isolate likely points at which experience-based learning might have occurred. For example, a subject choosing to listen as a moderate who receives no message might wonder why, and see that there could be a strategic advantage to listening as an extremist – thus discovering Bayesian reasoning. However, regression specifications noting subjects' exposure to such experiences, either period-to-period or with greater lag, all returned insignificant learning effects.

[20]Note that this is a conservative standard that cuts against the affirmative classifications of the Watsonian and the Overspeaking Bayesian types.

[21]For 15 of the 19 Overspeaking Watsonian/Overspeaking Bayesian subjects, and for all 3 of the Bayesian subjects, the observed choice pattern corresponds *exactly* to the expected behavior for the corresponding ideal types.

[22]A subject has an incentive to vote against her prior based on a foreign fragment only (1) when the

subject has 'B' active fragment under Listening Case (ABB); (2) when that subject chooses to listen; (3) when the counterpart agent with 'A' active fragment chooses to speak; and (4) when 'A' turns out not to be part of the subject's true number. Because, in Listening Case (ABB), the subject's true number is by definition more likely to be $\overline{AB}$ than $\overline{BC}$, and because, at least from the Bayesian perspective, it is out-of-equilibrium for the extreme agent to choose to speak in such a setting, all 4 conditions are met simultaneously only a small fraction of the time.

[23]Our post-experimental survey asked subjects to explain how they chose to speak vs listen; whether communication was helpful in deciding how to vote and if so, how; and how difficult they thought the problem posed was for them, and for others. We blind-coded subjects' responses into ideal types according to the type of reasoning they suggested. These codings yielded a very high correlation with actual behavior in the experiment, providing further support for our interpretation of the results. All 4 subjects classified as Bayesian [or Leaning] in the data gave survey responses consistent with Bayesian reasoning. Of the 20 subjects classified as Overspeaking Watsonian [or Leaning]/Overspeaking Bayesian, 14 gave survey responses consistent with Overspeaking Watsonian reasoning, 4 gave responses consistent with Overspeaking Bayesian reasoning, and 2 gave responses that defied straightforward classification.)

Table 1a. Weakly Dominant Deliberative Strategies for Bayesian Agents

Deliberative Setting	A active (Extreme)	B active (Moderate)	C active (Extreme)
ABB, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	Speak	Listen	-
ABB, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	**Listen**	Listen	-
ABC, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	Speak	Listen	**Listen**
ABC, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	**Listen**	Listen	Speak

Note: The weakly dominant deliberative strategies for agents with an 'A' or 'C' active fragment are indicated in plain text for the Speaking Case situations, and in boldface for the Listening Case situations.

Table 1b. Dominant Voting Strategies for Bayesian Agents with a 'B' (Moderate) Active Fragment, by Communications Received

Deliberative Setting	nothing	foreign fragment only	latent fragment
ABB, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	$\overline{BC}$	$\overline{BC}$	true number
ABB, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	$\overline{AB}$	$\overline{\mathbf{BC}}$	true number
ABC, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	$\overline{BC}$	$\overline{BC}$	true number
ABC, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	$\overline{AB}$	$\overline{AB}$	true number

Note: subjects with an 'A' ('C') active fragment always have a dominant strategy to vote for $\overline{AB}$ over $\overline{BC}$ ($\overline{BC}$ over $\overline{AB}$).

Table 2a. Aggregate Communications Behavior (Periods 13-30)

Deliberative Setting	A active (Extreme)	B active (Moderate)	C active (Extreme)
ABB, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	79.2% Speak (38/48)	99.0% Listen (95/96)	-
ABB, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	54.2% Speak (26/48)	97.9% Listen (94/96)	-
ABC, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	78.3% Speak (47/60)	96.7% Listen (58/60)	58.3% Speak (35/60)
ABC, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	66.7% Speak (40/60)	98.3% Listen (59/60)	73.3% Speak (44/60)

Table 2b. Aggregate Communications Behavior When Subjects Have 'A' or 'C' (Extreme) Active Fragments

.	Speak (S Case)	Listen (S Case)	Speak (L Case)	Listen (L Case)
Experimental Data				
All Periods	196 (71.0%)	80 (29.0%)	169 (61.2%)	107 (38.8%)
Periods 1-12	67 (62.0%)	41 (38.0%)	68 (63.0%)	40 (37.0%)
Periods 13-30	129 (76.8%)	39 (23.2%)	101 (60.1%)	67 (39.9%)
Theoretically Expected Behavior				
Bayesian	100%	0%	0%	100%

L Case - Listening Case; S Case - Speaking Case.

Table 2c. Aggregate Voting Behavior of Subjects with a 'B' (Moderate) Active Fragment (All Periods)

Deliberative Setting	nothing	foreign fragment only	latent fragment
ABB, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	91.7% $\overline{BC}$ (44/48)	91.8% $\overline{BC}$ (67/73)	92.7% true number (38/41)
ABB, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	94.6% $\overline{AB}$ (70/74)	56.3% $\overline{BC}$ (18/32)	100% true number (59/59)
ABC, $Pr(\overline{BC}\|B) > Pr(\overline{AB}\|B)$	100% $\overline{BC}$ (18/18)	95.5% $\overline{BC}$ (21/22)	96.4% true number (54/56)
ABC, $Pr(\overline{BC}\|B) < Pr(\overline{AB}\|B)$	100% $\overline{AB}$ (18/18)	85.7% $\overline{AB}$ (12/14)	90.6% true number (58/64)

Note: subjects with an 'A' active fragment voted for $\overline{AB}$, their dominant strategy, 96.7% of the time (348/360). Subjects with a 'C' active fragment voted for $\overline{BC}$, their dominant strategy, 96.4% of the time (185/192).

Table 3. Factors Affecting Deliberative Choice in Subjects with 'A' or 'C' (Extreme) Active Fragments (Periods 13-30)

Probit Regression with Robust Standard Errors

Dependent Variable: listen = 1 if subject listens, = 0 if subject speaks

$N = 336$; $PseudoR^2 = 0.0321$

listen	coefficient	robust SE	z	$P > \|z\|$
lnperiod	-.1352728	.3447588	-0.39	0.695
caselisten	.4685002**	.1472395	3.18	0.001
ABB (not ABC)	.0563819	.1722589	0.33	0.743
right-handed	.2296582	.1752874	1.31	0.190
xdist	.0100337	.0216525	0.46	0.643
xfrac	-.4254516	1.569162	-0.27	0.786
pdist	-1.562123	2.172081	-0.72	0.472
pfrac	.7623341	1.506065	0.51	0.613
constant	-.4361045	1.111846	-0.39	0.695

lnperiod is the natural log of the period number. **caselisten** is 1 for the Listening Case, 0 for the Speaking Case. **ABB (not ABC)** is 1 for ABB or CCD, 0 for ABC or BCD. **right-handed** is 1 for BCD or CCD, 0 for ABB or ABC. **xdist** is $|y_1 - y_2|$, the distance between the alternatives to be voted on. **xfrac** is $\frac{|y_1 - y_2|}{\overline{CD} - \overline{AB}}$, the distance between the alternatives to be voted on relative to the distance between the most extreme true numbers. **pdist** is the absolute value of the difference of the unconditional probabilities associated with the alternatives to be voted on. **pfrac** is **pdist** divided by the sum of the unconditional probabilities associated with the alternatives to be voted on.

Table 4. Individual Communications Behavior When Subjects Have 'A' or 'C' (Extreme) Active Fragments (Periods 13-30)

Subject No.	Speak (S Case)	Listen (S Case)	Speak (L Case)	Listen (L Case)	Classification
1	0	3	0	5	Dev
2	4	1	1	2	Unc#
3	6	0	3	3	OW/OB#
4	4	1	2	1	OW/OB
5	5	0	3	0	OW/OB*
6	3	0	0	5	B*
7	1	2	2	3	Dev*
8	6	0	0	6	B
9	2	4	2	4	Dev
10	1	5	1	5	Dev#
11	4	2	1	5	Lean B**
12	5	0	3	0	OW/OB#
13	2	4	6	0	Dev
14	3	0	5	0	OW/OB#
15	3	0	3	2	OW/OB
16	5	0	3	0	OW/OB*
17	2	1	3	2	Dev
18	5	0	3	0	OW/OB
19	3	0	5	0	OW/OB*
20	5	0	3	0	OW/OB*
21	5	1	5	1	OW/OB*
22	2	3	0	3	Dev
23	5	0	3	0	OW/OB###
24	3	0	5	0	OW/OB**
25	3	0	5	0	OW/OB*
26	6	0	3	3	OW/OB*
27	6	0	0	6	B*
28	6	0	6	0	OW/OB#
29	6	0	4	2	OW/OB*
30	5	0	1	2	Unc*
31	4	2	6	0	Lean OW/OB
32	1	2	4	1	Dev*
33	0	3	0	5	Dev###
34	2	3	2	1	Dev##
35	2	1	5	0	OW/OB
36	4	1	3	0	OW/OB#*
TOTALS	129 (76.8%)	39 (23.2%)	101 (60.1%)	67 (39.9%)	.

Dev - Deviant; Unc - Unclassified; B - Bayesian; OB- Overspeaking Bayesian; OW - Overspeaking Watsonian; Lean B - Leaning Bayesian; Lean OW/OB - Leaning OW/OB; L Case - Listening Case; S Case - Speaking Case.

The votes of those subjects who in any period of the experiment received only a foreign fragment while possessing active fragment B in the Speaking Case (ABB) are indicated next to their classification type. Each occurrence of (#) denotes one incorrect vote for the prior while each occurrence of (*) denotes one correct vote against the prior in such a situation.