Deliberation as Interactive Reasoning

William Minozzi[*] Michael A. Neblo[†] David A. Siegel[‡]

January 6, 2012

Abstract

Most formal models of deliberative democracy posit actors who want to guide a collective choice. This assumption stands in stark contrast to deliberative democratic theory as it was originally developed. The baseline non-formal model of this theory is the ideal speech situation, in which actors aim to understand and be understood, rather than to manipulate the outcome (Cohen 1989, Habermas 1990). Actors in this model are principally interested in the reasons each has for embracing a particular alternative. Deliberation is cast as a group hunt for sound, consensus rationales rather than as a game of strategic information transmission. We present and analyze a formal model of the ideal speech situation. Each actor is endowed with a set of inferences that she uses to guide her reasoning. During deliberation, actors can make assertions and disavow previous claims, query each other for reasons and challenge statements they disagree with. The goal is to reach consensus on a rationale for making a collective choice on the problem at hand given a limited amount of time. Using this baseline model, we characterize how deliberation changes with the composition of the deliberative body.

[*] Assistant Professor, Department of Political Science, 2137 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-247-7017, Email: william.minozzi@gmail.edu

[†] Assistant Professor, Department of Political Science, 2114 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-292-7839, Email: neblo.1@osu.edu

[‡] Assistant Professor, Department of Political Science, 541 Bellamy Building, Florida State University, Tallahassee, FL 32306 Phone: 850-945-0083, Email: dsiegel@fsu.edu

Many deliberative democrats think formal theory has little to offer in the analysis of deliberative phenomena. The main reason is that existing formal theories are based on an entirely different set of motivations than the ones deliberativists emphasize in their conceptions of the practice. For example, game theory is fundamentally the study of strategic interaction, whereas deliberation presumes that participants place at least some limits on responding to their own, private incentives. Indeed, in idealized cases, the focus is instead on seeking to understand and mutually accommodate each other via public and transparent discussion. Consequently, no matter how clearly game-theoretic depictions of political talk are rendered, many deliberative democrats believe there is not all that much to be gained by engaging a formal theory that can take no account of this difference. Game theoretic models might illuminate the limiting cases of deliberative failure, or augment our understanding of the "post-talk" phase of some deliberative practices. But they cannot shed much light on the core phenomena because conceiving of public reasoning in purely instrumental terms simply misses the heart of deliberation.

Formal theorists, for their part, argue that deliberativists present the veneer of well-grounded theory without fully warranting their claims or sufficiently accounting for incentives. The practice of writing down and analyzing a formal model forces one to answer hidden questions and explore unforeseen implications. For example, even though incentives play only a limited role in deliberative theory, systematically accounting for how and why participants would want to communicate might lead to counterintuitive conclusions. In short, there has been little productive interaction between these two fields, at least as would be recognized by scholars from the other side.

This impasse is lamentable but not inevitable. Non-cooperative game theory is not the only kind of formal theory. Many deliberativists recognize a role for social choice theory to supplement deliberative procedures. Habermas (1990), Cohen (1997), and others offer what even formal theorists would recognize as proto-formal theories, with readily identifiable actors and choices, albeit with unfamiliar and heretofore unformalized motivations. Staying close to these accounts, we develop a fully formal theory of deliberation as interactive reasoning, and argue that the result remains recognizable as a reasonable representation of the surprising overlap between the key elements in Habermas (1990), Cohen (1997), and Brandom (1994.). Because our model (like any plausible specification of deliberation) is complex enough to preclude analytical solutions, we analyze it using computational modeling (Kollman, Miller, and Page 1992; Bendor, Diermeier, and Ting 2003; Siegel 2009). Our formal theory is not only much better at capturing the conceptual heart of normative theories of deliberation, but it also provides a flexible baseline for moving away from ideal deliberation, setting the stage for rapprochement with more strategic conceptions of political discourse.

Absent such rapprochement, it is likely that each side will continue to press its valid complaint against the other without providing a way forward. Yet both sides face great costs from the standstill. The case for deliberative reform proceeds from two claims. First, as a theory, deliberative democracy has some very attractive normative properties (Habermas 1996; Gutmann and Thompson 2004). Second, in practice, deliberation tends to change things e.g., opinions, rationales, intensity, attitudes toward opposing views, etc. (Fishkin and Luskin 1999; Gastil and Dillard 1999). From these two premises, it may seem reasonable to infer that we should move toward implementing deliberative institutions. But there is a buried premise here. The conclusion does not follow unless we also assume that deliberation changes opinions primarily via mechanisms specified in the normative theories. Otherwise the argument gives us no warrant for believing that the changes are for the better. For, if the real sources of opinion change are morally inert, deliberation would, at best, waste social resources (Lupia 2002). And worse, if those sources include such mechanisms as social power, group conformity, etc., deliberation would magnify social inequality and pervert its own goals (Sanders 1997). Thus, we must carefully investigate the mechanisms of deliberative opinion change not only because the scientific questions raised are intrinsically interesting, but also because the normative argument for deliberative reform does not go through without it. The injunction to "first, do no harm" surely applies a fortiori to the body politic as well.

Moreover, even if we do decide that deliberative institutions deserve our support, we will want to know how to design them so as to further the normative goals of deliberation most effectively. Formalization is the next crucial step in isolating the mechanisms of deliberative opinion change.

When crafting a formal theory, many questions arise and demand attention. For example, we must model how a participant decides whether to ask a clarifying question, or to challenge her interlocutor with a counterargument. We must determine how a participant forms a new opinion or spies a previously unarticulated inference. While these may seem like mundane details, the way in which we model them animates what emerges. To formalize deliberation at all, we cannot gloss over such details and focus on seemingly more important topics. And this is a good thing. The mechanism by which better arguments have force emerges from a complex web of modeling decisions just like these. In this way,formalization aids in the difficult process of operationalizing and measuring key concepts like equality, sincerity, and motivation.

For deliberativists, linking causal mechanisms with an idea of how to measure important concepts provides much better leverage on the consequences of moving away from an ideal notion of deliberation and toward practical implementation. It is in this sense that a formal theory of deliberation is a baseline. The purpose of constructing such an ideal is not that we can hope to actually "achieve" it in any straight-forward way, but rather that it constitutes a standard against which we can judge our forays into the non-ideal world. A baseline formal theory allows us to formulate hypotheses about the likely effects of moving from ideal motivational, structural, and cognitive assumptions by changing one at a time. Articulating a baseline formal theory predicated on baseline assumptions and then comparing it with another that relaxes one of these constraints, say sincerity (Gutmann and Thompson 1996) or an orientation toward consensus (Mansbridge et al. 2010) generates hypotheses about the consequences of that change. This exploration may be expanded to include the interactive effects of relaxing multiple assumptions. Thus, a baseline formal theory of deliberation opens up conceptual space, providing potentially fruitful options and alerting us to the possibility of blind alleys.

The benefits of breaking the impasse between deliberativists and formal theorists seem to be better perceived by members of the latter group, whose previous attempts have roused skepticism from the members of the former. Landa and Meirowitz (2009) attempt at conciliation argues that a thorough account of incentives would actually help to build a better normative theory of deliberation on its own terms. Game-theoretic models are built upon what many would regard as an elegant, parsimonious representation of incentives, and come equipped with well-understood tools for analyzing the consequences of strategic interaction. But this comes at a price: game-theoretic models conceptualize reasons, a central feature of deliberative theory, in a way that fails to capture their normative and empirical functions.

Patty (2008) presents an alternative model of deliberation as collective choice in the space of reasons. In Patty's conception, arguments are paths through reasons, and each participant chooses whether to veto an argument based on the practical consequences of doing so. This model comes closer to capturing the idea that deliberation opens up conceptual space in its novel idealization of arguments. But the model also has a decidedly ends-oriented frame that is at odds with the primary theoretical assumption of deliberation as seeking understanding. Instead, Patty's model seems to be a model of pure sophistry.

In these models, and all other rational-choice treatments of deliberation, reasons are rendered in terms of their practical consequences. For example, in Landa and Meirowitz (2009), reasons *are* mere information, which is defined with respect to how it alters a participants beliefs about the consequences of reaching a certain conclusion (e.g., choosing a policy or reaching a judgment). Landa and Meirowitz (2009) write that, "In revealing correct, fuller, or simply better organized information, deliberation provides an opportunity for participants to arrive at more considered judgments themselves" (p. 427). The problem for participants in a game-theoretic deliberative setting is thus that information is either dispersed or processed incorrectly. This notion of information certainly models one important role played by reasons in deliberative theory, but it fails to capture their main function.

In any adequate model of deliberation, a reason should be a statement that is able to stand as both a premise and a conclusion in an inference.[1] In game-theoretic models by contrast, a piece of information cannot be questioned or challenged; things which can be inferred from it and things that permit it to be inferred are absent. A piece of information may be countered by another piece of information that points in the opposite direction, but there is no way for two participants, through communication, to reconcile the two. This omission is both more fundamental and more problematic than game theorys assumption that all participants are purely strategic. Deliberation is not just an opportunity to learn things others know or to better organize isolated units of information, but to more fully articulate a public justification for actions on matters of common concern.

What Should a Formal Theory of Deliberation Do?

We have argued above that nearly all deliberative theorists reject as incomplete the leading formal-theoretic replacements for reasoningstrategic revelation of information and incentive-driven restriction of acceptable argumentsbut we have not described in detail what might be acceptable. Before presenting our formal theory, we recapitulate the processes and actions that several early and influential deliberativists envision as a baseline standard of the practice of deliberation. First, deliberativists understand the ideal motivations of players as being oriented toward mutual understanding and the hope of coming to some level of agreement (Cohen 1997; Habermas 1990). It is does not suffice to imbue participants with common preferences in the rational-choice theoretic sense. Rather, ideally motivated participants want to learn the reasons for why they agree or disagree; they must be driven not only by a search for their personal notion of the best policy, but by a search for the reasons that would warrant them and their fellow citizens in believing a policy to be the best. Unlike formal theories of deliberation built exclusively on the firmament of instrumentally rational

[1] Our notion of reasons builds on the foundation established by Brandom (1994.), Cohen (1997), and Habermas (1990).

choice, participants regard how they reach deliberative outcomes as internally related to having reasons for the desirability of the outcomes themselves.

This focus on motivation immediately draws ones attention to the process of interactions, which we argue most deliberativists would consider a second essential element of any formal theory that purports to model deliberation. Here, we distinguish persuasion on the merits from mere rhetorical effectiveness. In the ideal case, persuasion on the merits requires that participants communicate with each other through vulnerable reasons. A reason is vulnerable in the sense that each other participant can call the validity of the reason into question, and can either accept or reject it. For the "unforced force of the better argument," which is the only means by which one can compel others within an ideal deliberative setting, to make sense within a formal theory, participants must be able to evaluate the goodness of reasons. In formal theories of deliberation based on instrumentally rational choice, a reason provided by one participant is only persuasive to another if the provider and receiver have similar preferences and the provider has some authority. Such theories cannot make sense of the internal persuasive force of the reasons themselves. In contrast, while there is some role for authority in an ideal deliberative setting, the exchange of reasons is closer to a joint exploration of the inferential properties of what beliefs we have in common than it is to a measure of how similar participants preferences are.

A Theory of Deliberation as Interactive Reasoning

We describe the model twice: first informally, and second in more techical terms. In the abstract environment we construct and analyze, there are two fundamental objects: participants and statements.[2] Each participant is fully characterized by a description of her cognitive structure, and the means by which she forms discursive priorities, both of which we describe in detail below. A cognitive structure is composed of opinions, each participants discursive priorities list encompasses speech acts, and both of opinions and speech

[2] Throughout the presentation of the theory, we use italics to present a term for which we provide a formal definition.

acts are defined in terms of statements. There are two varieties of statement: simple and complex. Simple statements are the basis for entry into deliberation. For example, a simple statement would be “That building is brown.” In contrast, a complex statement consists of other (possibly simple) statements and one or more of the operations “and”, “or”, “not”, and “because”. If a, b, and c represent (simple or complex) statements then “a because b”, “b and c”, and “not c” are both complex statements. Operations can be iterated to form the basis for more complicated opinions, such as “both a and b, because c”.

Building upon the fundamental idea of a statement, a participant forms an opinion that capture her beliefs about that statement, including how salient it is, how confident she is that the statement is valid, and a measure of how well her opinion on the subject coheres with her other opinions. A participant may move toward entering the deliberative fray by adding a speech act based on a statement to her discursive priorities list. The available types of speech acts are “assert”, “disavow”, “challenge”, “query-why”, and “query-whether”. Sample speech acts would be “I assert a because b” or “I disavow my prior assertion not c”. Each speech act on a participants list is prioritized by its urgency, which increases based on how germane it is to deliberation and what its potential impact for the overall coherence of the participants cognitive structure.

Each round of deliberation begins as a participant makes a speech act. All of the participants, including the speaker, then dwell on the speech act for a moment, and consider its ramifications for their other opinions and discursive priorities. Each participant then decides whether to add herself to a queue of future speakers, and the process is repeated. In the next two subsections, we present a detailed account of our participants cognitive structures and a round of deliberation.

Cognitive Structures

The cognitive structure of each participant consists of three objects: a web of beliefs, a set of opinions, and a discursive priorities list. The first is represents long-term memory;

the second, short-term memory. The last is a mental tally of speech acts the participant considers making later in deliberation. We elaborate on each of these objects in turn.

Before deliberation commences, participants have some preconceived notions about the subject of discussion. Formally, each is endowed with a web of beliefs, a network among simple statements that resembles long-term memory. Each participant associates with each simple statement an authority score, which is essentially the credibility the participant initially assigns to it. For example, a participant might assign a simple statement like “That building is brown” different authority scores depending on whether she was colorblind. A positive authority score means a participant finds a statement to be credible; negative scores are associated with statements the participant doubts. If the authority of a statement is 0, the participant has no preconceived notions about that statement.

Along with these authority scores, each participants web of beliefs encompasses inferences she makes between simple statements. The inferential link-weight from one simple statement to another is a number that measures the extent to which the participant believes the second to be a consequence of the first. These are directed links, which means that the inferential link-weight from a to b is potentially (though not necessarily) different from the inferential link-weight from b to a. For example, the inference from “He stabbed the victim” to “He held a knife” may have higher weight than the inference from “He held a knife” to “He stabbed the victim”. As with authority scores, inferential link-weights are positive for inferences the participant recognizes as good and negative for inferences the participant believes to be flawed.

As deliberation begins, each participant focuses her attention and cognitive resources on the issues at hand. In our theoretical environment, this means developing a set of opinions about other participants and about statements. First, each participant keeps score on her interlocutors with a reliability score, which increases when the interlocutor says something the participant ends up agreeing with, and decreases otherwise. Agreement is conceptualized in terms of a participants opinion on a statement. A participants opinion about a statement (which could be simple or complex) is composed of three numeric values: a confidence score, a salience score, and a coherence score. Like authority scores from the web of beliefs, confidence scores reflect beliefs. Unlike authority, confidence is assumed to be malleable. One effect of deliberation is the revision of confidence in light of reasons provided by others. Along with confidence, a participants opinion about a statement includes a measure of how salient that statement is. The salience of a statement increases as it is brought up in deliberation, and decays when the locus of discussion shifts.

Each participant also perceives the coherence of an opinion within her broader set of opinions. Coherence measures how well an opinion matches other opinions, taking into account their salience and confidence scores. Importantly, it is here for the first time that something like logic enters the picture. To determine the coherence of an opinion, a participant measures the compatibility of that opinion with each other opinion she currently has. The compatibility of a pair of statements is either 1, 0, or -1. A compatibility score of 1 means that the two statements concern a similar subject and are not logically exclusive. For example, the two statements "a because b" and b have a compatibility score of 1, the statements "a and b" and "not b" have a score of -1, and the statements "a because b" and c have a score of 0.[3] The coherence of a particular opinion is then the sum of the confidence the participant has in each other opinion, weighted by that opinions salience score, and signed by the compatibility of the pair of opinions. Thus, if two opinions do not have anything in common (i.e., compatibility is 0), or if neither is relevant to the conversation (i.e., salience scores are 0), or if the participant is agnostic about both (i.e., confidence scores are 0), they do not affect the coherence associated with each other.

Along with her web of beliefs and set of opinions, a participants cognitive structure includes a discursive priorities list. This list includes the speech acts that the participant wants to make, along with measures of the germaneness, potential impact, and urgency of

[3] Strictly speaking, this last example assumes that c is not further decomposable into substatements that render it either compatible or incompatible with "a because b". To operationalize this step, we have enumerated a list of pairs of statements which are taken to be incompatible.

those acts. Items on the list include, for example, “I assert that a because b”. Associated with each act on the list is an evaluation of how germane the act is to the conversation. Thus, the act in the example is more germane if someone recently executed a speech act involving a or b. An act has a high potential impact score if adding it to deliberation has the potential to change the overall coherence of the participants set of opinions. This change might be positive, or it might be negative. Returning to the example, if deliberation were to focus on consequences of a, each participants opinions would be peppered with many statements predicated on a. In this case, the speech act “I challenge a because b” would have high potential impact. Urgency is then a combination of germaneness and potential impact of an act, and participants give more priority to speech acts with higher scores on each.

At this point, we have described the cognitive structure of a participant: her web of beliefs, her set of opinions, and her discursive priorities list. Within this framework, deliberation is composed of participants making speech acts and examining the cognitive consequences of each others utterances. The particular process by which this occurs is the subject of the next subsection.

A Round of Deliberation

For illustration, we focus first on a round of deliberation in which a participant asserts a statement. The other speech act typesincluding disavowal, challenge, query-why, and query-whetherare then presented as variations on this paradigmatic type. Suppose a round begins with an assertion. First, this act, as well as the speaker who made it, is recorded in a running history of what has transpired. Then, all participants, including the speaker, evaluate the consequences of the assertion for her set of opinions. At the end of these independent, intra-participant cognitive processes, each participant updates her discursive priorities list, and decides whether to add herself to the queue of future speakers. Finally, the next speaker is recognized to speak, and the process repeats.

Given a particular assertion, the cognitive process of each participant unfolds as follows. Other speech acts are largely similar. First, the participant assigns the salience of the asserted statement increases, while the salience of all other statements in her set of opinions decreases. She then revises her confidence in the asserted statement based on the previous reliability score she assigns to the speaker. Simultaneously, the participant revises her reliability score of the speaker based on the confidence score she previously assigned to the asserted statement. She then updates her perception of how coherent the newly asserted statement is within her set of opinions.

The participant then repeats this process, but, rather than dwelling on the originally asserted statement, she focuses on a series of related statements. A statement is more likely to be selected as the focus of internal cognition if it is more salient. At this point there is also the potential for creative thought, which occurs when a novel statement occurs to the participant. Internal cognition continues probabilistically and eventually gives way to the next step, in which participants make amendments to their discursive priorities lists and prioritize their most urgent speech acts.

Other speech acts are most easily presented in terms of how they differ from the paradigmatic act of assertion. Disavowal mirrors assertion. After disavowing a statement, a participant can no longer be held to account for its assertion; she can no longer be the subject of queries demanding reasons to believe the statement or challenges to the statements validity. During the remainder of the round, participants examine the consequences of the negation of the statement that was disavowed.

The two sorts of queries differ from assertion and disavowal in that they single out another participant and demand a response from her. If a participant utters a speech act using query-why, she must specify a second participant and something that this latter speaker has already asserted. The second speaker is moved to the top of the queue and must assert a reason for her assertion. For example, if one speaker queries as to why a second asserted the statement a, the second speaker must utter something of the form “I assert a because b”. Similarly, if a speaker speaks using query-whether, she must specify a second participant and a statement, and the second speaker is obligated to offer a verdict on that statement. For example, if one speaker queries as to whether a second believes a statement a, the second speaker must utter either “I assert a” or “I assert not a”.

Finally, challenge requires a speaker to target another participant, a statement the second has asserted, and a reason to challenge that statement. Challenge is like the two sorts of query in that it moves its target to the front of the queue. It also functions as an assertion, focusing all participants attention on the validity of the proffered reason for the challenge. The targeted participant must then either disavow her previous assertion of the challenged statement or provide another reason, this one designed to cast doubt on the original reason offered by the challenger.

Motivation, Creativity, and Sincerity

Before presenting the formal model, it is worth pausing to dwell on what motivates participants, how participants “think”, and what participants are allowed to say. Participants are motivated to say things that are germane and have large potential impacts. The first criterion prioritizes speech acts that are related to the topic at hand, without absolutely excluding acts that concern more cursorily related subjects. The second prioritizes utterances that change the overall coherence of a participants set of opinions, positively or negatively, thus providing pressure to search through conceptual space and more fully articulate the inferential consequences of ones assertions. Importantly, neither of these motivations excludes any speech act from deliberation. Instead, it is simply more likely that a participant will make a particular speech act if it is more germane and potentially effective.

In a similar manner, there are chances throughout each round of internal cognition for participants to generate novel opinions. A novel opinion is based on a statement that is not currently within the participants set of opinions. To evaluate this new statement, the participant refers to her web of beliefs and assigns the statement a confidence score based on what she finds. The new statements coherence is evaluated, and it becomes fair game for future speech acts.

Although we have emphasized what participants can do in the model, it is also important to demarcate what they cannot do. Principally, participants are required to have some degree of confidence in what they assert. This is, in effect, a sincerity requirement.

The Formal Model

Now consider a more formal treatment of this model. There are *participants* P and a *set of simple statements* X. Throughout, p refers to a generic participant who has made a speech act; let P_{-p} denote the set of participants excluding p. Let the *set of statements* be S. The *basic operations* are mappings $F = \{\mathit{not}(\cdot), \mathit{and}(\cdot,\cdot), \mathit{or}(\cdot,\cdot), \mathit{because}(\cdot,\cdot)\}$.[4] The set of statements can thus be written as the union of recursive sets $S = \cup_j S_j$, where $S_j = \{s = f(t,u) | f \in F \text{ and } t,u \in S_{j-1}\}$ and $S_0 = X$.

Building on statements, participants communicate with each other sequentially in rounds of deliberation via *speech acts* D, which are recorded in *history* H. For each round of deliberation, the history is a pair (p,d) of a participant who made a speech act. A speech act begins with a *speech type*, each of which has different consequences. The speech types are $G = \{\mathit{assert}, \mathit{disavow}, \mathit{query\text{-}whether}, \mathit{query\text{-}why}, \mathit{challenge}\}$. Let $H_p(g)$ be the subset of histories in which participant p made a speech act of type g.

Each speech act depends on different pieces of information and has different consequences for future speech acts. The *assert* type is simply the assertion of statement $s \in S$, which is then considered by the other participants. Its counterpart *disavow* takes a statement $s \in H_p(\mathit{assert})$ that the participant previously asserted and puts it in doubt. The types *quey-whether* and *query-why* both identify a second participant $q \in P_{-p}$ and a statement $s \in H_q(\mathit{assert})$ that she previously asserted. Each prompts the second participant to assert a response. Finally, *challenge* also identifies a second participant $q \in P_{-p}$ and a statement

[4] The operation $\mathit{because}(s,t)$ has the meaning "s is true because of t."

$s \in H_q(assert)$ that q previously asserted. But *challenge* also requires a reason in the form of a second statement $t \in S$. Thus, *challenge* essentially combines three assertions: $assert(not(s))$, $assert(t)$, and $assert(because(not(s), t))$. Given these speech types, the set of speech acts can be defined as $D = \{d = g(t, u, q) | g \in G$ and $t, u \in S$ and $q \in P\}$.

Each participant p has a *cognitive structure* constituted by several components. First, a participant's *web of beliefs* is a pair of networks that represents their ideas about causal relationships among simple statements. Importantly, we do not hypostatize one true causal network; there is room for disagreement. Instead, causality is modeled as a coherence network. There are two main parts to this network. First, the *inferential link-length* between any two simple statements for participant p is $\lambda_p(s, t) \in [0, 1]$, which resembles the speech act $because(s, t)$. Second, the sense data available to participants. These take the form of *noninferential authority* scores $\theta_p(s) \in [-1, 1]$ for all $s \in X$.

Each participant p also has a set of opinions composed of four different scores. The first three scores pertain to a statement s. *Confidence* $\alpha_p(s) \in [-1, 1]$ is a measure of how much the participant agrees with statement s. *Coherence* $\kappa_p(s) \in [-1, 1]$ is a measure of how well s fits into p's web of beliefs. And *salience* $\sigma_p \in [-1, 1]$ measures how relevant s is to the matters that have been discussed recently. The fourth score, *reliability*, is a measure of how much p trusts another participant $q \in P_{-p}$, and it is denoted $\rho_p(q) \in [-1, 1]$. Below we discuss the interaction of opinions and webs of belief.

Cognition in this model is formally defined by several functions. For any participant p, two statements s and t are said to be *connected* if they are both in S_p and, for example, $s = and(s, t)$ or $s = not(because(s, t))$. Thus, define $connect_p(s, t)$ as an indicator function that equals 1 if and only if both statements contain at least one simple statement in common. In contrast, two statements are in *contradiction* if one depends on a substatement and the other depends on *not* that substatement. Thus, $contradict(s, t)$ is an indicator that equals 1 iff s and t include substatement s'' and $not(s'')$. Importantly, *contradict* is not participant-specific. Third, the function $substance(d)$ identifies the substantive statements underlying particular speech acts. For example, $substance(assert(s)) = \{s\}$. The *substance* is also not participant-specific. Finally, there is a function $agree_p(s,t)$ to measure whether the conjunction of two statements is logically coherent. Thus, $agree_p(s,t) = (1 - 2 \cdot contradict(s,t)) \cdot connect_p(s,t)$.

Participants choose how and when to speak by keeping a *discursive priorities list* L_p S. The elements of the lists are speech acts, and each is associated with three measures of its value as an utterance. For a speech act $d \in D$, these scores are *germaneness* $\gamma_p(d) \in [0,1]$, *potential impact* $\pi_p \in [-1,1]$, and *urgency* $u_p \in [0,1]$. To illustrate the list, consider a round of deliberation in which a participant makes an assertion, since all the other speech types are based on *assert*. In brief, each round has several stages (1) a participant p makes the speech act $assert(s)$, (2) the element $(p, assert(s))$ is added to the history H, (3) each other participant q incorporates $(p, assert(s))$ into their cognitive structures, evaluating it as they will, (4), all participants including p reformulate their discursive priorities lists, (5) any participant can add herself to the discursive queue, and (6) a new speaker is drawn randomly from the queue.

Before moving on, two steps of deliberation require more detailed description. The first is the essential deliberative step, number (3) from the list in the last paragraph. This internal cogitation step has several substeps. First, when participant q incorporates $assert(s)$ her cognitive structure, she proceeds as follows. First, she evaluates the *salience* of s. If s is not a new addition to S_q, then she revises salience $\sigma_p(s)$ upward, so its new value is equal to $\sigma_p(s) + \beta_\sigma$. At the same time, all other statements decrease in salience, so $\sigma_p(t)$ decreases for all $t \in S_q \diagdown \{s\}$ to the new value $(1 - \delta_\sigma)\sigma_p$. Similar steps are taken for both confidence for statements and reliability for participants. Next, coherence κ_q is revised according to the following rule:

$$\kappa_q(s) = \Sigma_{t \in S_q} \sigma_q(t) \cdot \alpha_q(t) \cdot agree_q(s,t)$$

Thus, coherence changes according to the salience of other connected statements with which s agrees or disagrees. Finally, the participant thinks about what she has learned. This part of the cogitation process is a loop that essentially replicates the last three steps, randomly choosing related statements, evaluating their salience, confidence, reliability and coherence, thus updating webs of belief.[5]

The last step of deliberation to discuss is the updating of discursive priorities lists. To update these lists, a participant focuses on the most recent speech acts, where d^t is the speech act from round $t = \tau$, and considers the acts that are currently on her list $d \in L_p$. Each of these acts is evaluated for germaneness, potential impact, and urgency. For speech act d, germaneness is given by

$$\gamma_p(d) = \sum_{t=0}^{\tau-1} \delta^{\tau-t} \left(\frac{\sum_{s' \in substance(d^{\tau-t})} \sum_{s \in substance(d)} connect_p(s, s')}{\#(substance(d^{\tau-t}))\#(substance(d))} \right)$$

Thus, a speech act is more germane if it has substance that is connected with other recent speech acts. Next, potential impact is based on the difference between current coherence in S_p and the conjectured coherence that would follow if p uttered d. The *coherence* of S_p (as opposed to that of a single statement s) is a weighted average of the coherence of all the statements in S_p, weighted by salience.

$$\pi_p(d) = \frac{\sum_{s \in S_p} \sigma_p(s)\widetilde{\kappa}_p(s)}{\sum_{s \in S_p} \sigma_p(s)} - \frac{\sum_{s \in S_p} \sigma_p(s)\kappa_p(s)}{\sum_{s \in S_p} \sigma_p(s)}$$

Conjectured coherence is $\widetilde{\kappa}_q\left(s\right)$ is given by

$$\widetilde{\kappa}_p(s) = \Sigma_{s' \in S_p \cup substance(d)} \sigma_p(s') \cdot \alpha_p(s') \cdot agree_p(s, s')$$

Finally, p evaluates the urgency of d. Urgency is simply Cobb-Douglas utility function of germaneness and potential impact $u_p(d) = aln\gamma_p(d) + (1-a)\ln \pi_p(d)$. Given this model, we now offer a brief description of some preliminary results.

[5] A slightly amended process occurs when a new statement is encountered, the details of which we omit here.

A Preliminary Look at What Emerges

Above we outlined a verbal description of the assumptions and processes in our theory of deliberation. In this section, we walk through two sample runs of a computer program that operationalizes these rules. To be clear, what follows should be taken as evidence that a theory like the one we describe can be written down and analyzed, rather than as anything approaching a complete analysis of such a theory.

In the examples we discuss, we chose parameter values with the goal of keeping the results simple. Thus, there are only three participants and seven simple statements in the example. We set the cognitive capacity parameters to low levels, and speech acts are followed by only a single round of internal cognition. The likelihood a participant will spawn a novel opinion is also low. Participants enter deliberation with webs of belief that are randomly generated (i.e., there is no correlation in webs of belief driven by facets of either the objective or social worlds), with the sole restriction that each has a salient opinion on a central focus of deliberation. For illustrative purposes, we have named this statement "guilty".

While examination of a single run of this program cannot reveal much about the generic properties of the theory, it can serve as a proof of concept, and illustrate aspects of the kinds of output that the model can generate. First, we ran the program for 10 rounds for a particular random draw of starting webs of belief. Recall that an authority score is positive if a participant believes the associated statement to be true, and negative if false. In this particular example, Participant 1 had a weak belief that "guilty" was true, with an associated authority score of +0.1; Participant 2 had a strong belief that "guilty" was true, with authority +0.9; and Participant 3 had a middling belief that guilty was false, with authority -0.4. Each also had some initial discursive priorities: chiefly, to discover what the others believed. The first speech act comes as Participant 1 queries 3 about her thoughts on the guilty question. Participant 3 then replies by asserting "not guilty". Next, 2 asks 1 her thoughts, and 1 replies by asserting "guilty". Thus deliberation begins as the participants explore who stands where (albeit in the coarsest of terms). To get a feel for what plays out in a longer term, we ran another random draw of the program for 100 rounds. Here, we focus on belief change over the course of deliberation, and the means by which this change occurred. At the beginning of this run, 1 had a strong belief that “guilty” was true, with authority +0.9; 2 had a somewhat weaker notion that “guilty” was false, with authority -0.5; and 3 had a slightly stronger belief that “guilty” was false, with authority -0.7. By the end of round 100, these beliefs had polarized, with 1 moving from +0.9 to +1.0, 2 moving from -0.5 to -1.0, and 3 moving from -0.7 to -1.0. In addition to more solid opinions, participants built up some simple reasons for their beliefs.

For example, participant 2 based her belief that “guilty” was false on a reason with the label cshe assigned c the confidence score +0.6 and the inference “not guilty because c” confidence +0.9. This result indicates that even such a “thin” run of the model captures the idea of reason-based deliberation, if only in a crude way for now.

Further Research

Although much remains to do in both the basic construction of this theory and analysis of it, once we have completed a baseline formal theory of deliberation worthy of the name, there is no shortage of directions for future research. First, however, we describe what a fuller analysis of this theory would look like. There are many parameters in the theory, including the numbers of participants and simple statements, several values that capture cognitive capacity, and the length of deliberation. The initial webs of belief with which participants enter into deliberation might be more or less congruent; the deliberative body might be bifurcated or factionalized into many subgroups with internally consistent but externally discordant webs of belief. Moreover, we randomly generate many specific elements of each example, including the occurrence of spontaneous, creative sparks that manifest in opinions on novel subjects. To fully explore such a complex environment, we must run many examples so that we learn which of the phenomena that emerge are rare and which are commonplace.

We are interested in many questions that such an investigation can answer. How many rounds does it take for stable opinions to emerge on important subjects (or under what conditions do stable opinions emerge at all)? Does this change depending on the number of participants, or the degree of their diversity of beliefs, or their cognitive capabilities? Moreover, we could consider the role of the implicit homogeneity assumption that all players have the same cognitive capacities and priorities over speech acts.

Once we have established the properties of a baseline theory, we can stride out into conceptual space and explore how departing from this ideal alters our theoretical expectations. One clear avenue to consider is the role of motivation and personality types. Not only could we include strategic participants, who want to reach what they consider to be the best outcome regardless of how they do so, we could also include argumentative gadflies who simply relish challenging others assertions. It would be particularly interesting to learn whether participants can use their measures of reliability, which they use to keep score on each other, to isolate such spoilers, or whether this deliberation in this environment is vulnerable to non-motivationally ideal behavior. Other participants might have fixed opinions that are resistant to revision, such as rigid religious or ethical beliefs, and whether deliberation in this setting can establish accepted public reasons in spite of such deep pluralism. No doubt each reader will have his or her particular concern about what our current statement of the theory is missing, or how we implement a given element of deliberative theory. Indeed, we have many ideas along these lines ourselves, and hasten to admit that this project is in its infancy. But by bridging the basic conceptual gulf that separated game-theoretic and normative conceptions of deliberation, we hope to have provided a formal language and platform for sharpening and advancing theoretical, normative, and empirical debates about one of the most promising developments in political theory and practice of the last half century.

References

Bendor, Jonathan, Daniel Diermeier, and Michael M. Ting. 2003. "A Behavioral Model of Turnout." *American Political Science Review* 97(2): 261–280.

Brandom, Robert. 1994. *Making It Explicit.* Cambridge: Harvard University Press.

Cohen, Joshua. 1997. *Deliberative Democracy: Essays on Reason and Politics.* Cambridge: MIT Press chapter Deliberation and Democratic Legitimacy, pp. 67–92.

Fishkin, James S., and Robert C. Luskin. 1999. *The poll with a human face: The National Issues Convention experiment in political communication.* Routledge chapter Bringing deliberation to the democratic dialogue, pp. 3–38.

Gastil, John, and James P. Dillard. 1999. "Increasing Political Sophistication through Public Deliberation." *Political Communication* 16(3): 3–23.

Gutmann, Amy, and Dennis F. Thompson. 2004. *Why Deliberative Democracy?* Princeton:: Princeton University Press.

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

Habermas, Jurgen. 1990. "Discourse Ethics: Notes on a Program of Philosophical Justification." In *Moral Consciousness and Communicative Action.* MIT Press.

Habermas, Jürgen. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy.* Cambridge, MA: MIT Press.

Kollman, Ken, John H. Miller, and Scott Page. 1992. "Adaptive Parties in Spatial Elections." *American Political Science Review* 86(4): 929–937.

Landa, Dimitri, and Adam Meirowitz. 2009. "Game Theory, Information, and Deliberative Democracy." *American Journal of Political Science* 53(2): 427–444.

Lupia, Arthur. 2002. "Deliberation disconnected: What it takes to improve civic competence." *Law and Contemporary Problems* 65(3): 133–150.

Mansbridge, Jane, James Bohman, Simone Chambers, David Estlund, Andreas Follesdal, Archon Fung, Cristina Manin Lafont, and Jose Luis Bernard Marti. 2010. "The Place of Self-Interest and the Role of Power in Deliberative Democracy." *Journal of Political Philosophy* 18(1): 64–100.

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

Sanders, Lynn M. 1997. "Against Deliberation." *Political Theory* 25(3): 34776.

Siegel, David A. 2009. "Social Networks and Collective Action." *American Journal of Political Science* 53(1): 122–138.