Creative Synthesis:
A Model of Reflective Equilibrium and Ideology Formation*

Hans Noel
Robert Wood Johnson Scholar in Health Policy Research
University of Michigan
hansnoel@umich.edu

109 Observatory
SPH II / M2242
Ann Arbor, MI 48103

April 7, 2010

Abstract

This paper offers a formal model of ideological formation that combines psychological predispositions and rational self-interest. I argue that by modeling the way in which political thinkers reason from first principles, and how they fail to ignore their own psychological and interest-based biases, we can explain ideological development. A model of long coalitions (Bawn, 1999) provides a structure for people's interests and their psychological traits, and a model of reason (Rawls, 2001) provides a method of combing those interests and traits into an ideology.

[*] Paper prepared for presentation at the annal meeting of the Midwest Political Science Association, Chicago, Ill., April 2010. I would like to thank Kathy Bawn, Sean Gailmard, Ken Kollman, Jeff Lewis, Skip Lupia, Chloé Yelena Miller, John Patty, Ken Schotts, Tom Schwartz, John Zaller and seminar participants at the University of Michigan for useful comments on earlier versions of this project.

> "[T]he shaping of belief systems of any range into apparently logical wholes that are credible to large numbers of people is an act of creative synthesis characteristic of only a miniscule proportion of any population."
>
> Philip Converse, *The Nature of Belief Systems in Mass Publics,* 1964, p. 211

Introduction

Words, we are told, have power. Pens are mightier than swords, and ideas are both bulletproof and dangerous. But this endorsement of ideas and the words that articulate them is as vague as it is enthusiastic.

This paper is about a particular way in which ideas matter. Whatever other power they may have, ideas can win because they get help. In the model the follows, ideas get the help of other ideas, forming coalitions of ideas that we call ideologies.

This is a model of ideology that integrates our best ideas about where ideology comes from, without reducing ideology to a caricature. A rich literature dating back to at least Adorno argues that ideology has psychological roots in personality traits. Meanwhile, another tradition dating back to at least Marx says that ideology is a rationalization of interests. Finally, for many, ideology is not just something that scholars study but something that they sometimes do. Ideological thinkers at least aspire to develop good theories of social order and justice, and to persuade others that they are right. Even if scholars believe ideologues are limited by their personalities or their self-interest, there is likely some consequence of their efforts to rise above those limits. The result is an ideology that functions like a coalition, linking ideas with other ideas.

The paper proceeds in six parts. In Section 1, I provide a formal environment for understanding ideology. Section 2 discusses ideology's psychological-, interest-, and principle-based roots. Section 3 explores a model of philosophical reasoning advocated by John Rawls. While his notion of reflective equilibrium, may have flaws as a normative prescription, it probably does describe the work of ideologues. It thus allows us to introduce philosophical reason to the origins of ideology. Section 4 discusses a formal model of the process of underlying Reflective Equilibrium. Section 5 presents results from simulations of that model. Section 6 discusses potential extensions.

1 Ideology as Constraint

I proceed from a sparse definition of ideology. Gerring (1997, p.980) surveys the literature and identifies one core element of ideology, coherence or constraint: “Ideology, at the very least, refers to a set of idea-elements that are bound together, that belong to one another in a non-random fashion.”

This notion of constraint is generally operationalized as a set of policy positions that go together, often defining an ideological dimension Knight (2006). So, if we are to set aside *why* they are bound together for a moment, ideology is something that bundles issue positions. Ideology is about what goes with what. The result might be an ideological dimension, familiar both theoretically (Downs, 1957; Black, 1958; Enelow and Hinich, 1989; Hinich and Munger, 1994) and empirically (Poole and Rosenthal, 1997; Poole, 2000).

This ideology is generally viewed as a single dimension, although it might naturally break into multiple sub-dimensions —social, economic, foreign policy, racial. Part of the notion of “what goes with what” is exactly that social conservatism goes with economic conservatism, for example, even though they need not go together logically.

The question of why things might go together looms large in such a sparse definition. We will return to that question in section 2. For now, ideology is constraint. But constraint can mean many things. Scholars generally operationalize ideology as constraint through a set of questions (or votes, or items) which actors answer similarly. If you favor abortion rights, you should favor labor unions and oppose the War in Iraq. If you oppose abortion, you take the opposite position on those issues. Our first step, then, is to get more rigorous with our definition of ideology and of constraint.

1.1 A Model of Constraint

Begin with a set of issues and actors with positions on those issues. However, this is not a theory of general ideological adoption, but of ideological generation. So, following Converse, we restrict the actors to those who are part of the “miniscule proportion of the population” that engages in the “creative synthesis” of ideology:

> The second source of social constraint [the first is interests, as in section 2.2] lies in two simple facts about the creation and diffusion of belief systems: First, the shaping of belief systems of any range into apparently logical wholes that are credible to large numbers of people is an act of creative synthesis characteristics of only a miniscule proportion of any population. Second, to the extent that multiple idea-elements of a belief system are socially diffused from such creative sources, they tend to be diffused as "packages," which consumers come to see as "natural" wholes, for they are presented in such terms ("If you believe this, then you will also believe that, for it follows in such-and-such ways"). 1964, p. 211

I accept without objection the mechanisms laid out elsewhere (see especially Zaller, 1992) for how those elite opinions reach ordinary voters. Much of the study of ideology has focused on the attitudes once they have diffused into the electorate. The attitudes of voters are important, but if we believe the source of their attitudes is elite discourse, then we need to model that discourse. This is determined by these elite thinkers, whom I argue are making a good faith effort to decide what is best. We might call this group "an intellectual elite," or "a discourse" or "the chattering classes." I will use the term "intelligentsia." They begin the process.

The model includes:

- X is a set of k binary issues.
- P is a set of n intellectual actors.
- $y_{ij} \in \{0, 1\}$ is the i^{th} actor's "position" on the j^{th} issue.
- $\mathbf{Y}$ is an n by k matrix of issue positions.

Constraint means that one's position on one issue determines or influences one's position on another. But if we are open to any sort of reason behind that constraint, then *any* pattern could be an ideology. Who are we to say that one bundling is more legitimate than another? For example, we might be inclined to say that a person who favors collective bargaining rights for labor unions but not a mandated minimum wage is not ideological, because those two policies are logically and politically connected. But a person who opposes the death penalty but favors abortion rights might seem to violate an equally valid logic.

In practice, scholars typically determine which pattern of constraint is valid by looking across the rest of the data. This is literally what NOMINATE and similar scaling methods do. Psychologists sometimes ask questions about whether or not actors seem to have a grasp of the ideological concepts that are used to organize issues, but those concepts tend to be chosen from those that are widely used by liberals and conservatives. And for our purposes, there is little wrong with this. "Liberalism" and "conservatism" (and some other, similar bundles, like "libertarianism") are the empirically observed ideologies that matter in politics. The question is what causes or is capable of causing them.

And so, we can operationalize an ideology in this way:

- β is a **belief system**, or a set of issue positions that are meant to be internally consistent.

A belief system becomes an **ideology** when it is shared by many actors. We could then ask of an intelligentsia, how many ideologies are there, and how different from one another are they? It is useful to think of a matrix of issue positions, resembling a roll call matrix typically used in scaling techniques. In that case, we can identify the ideologies that seem to anchor the system and give it structure.

Let us take this notion to the basic framework here of preferences over different issues suggested by the matrix $\mathbf{Y}$ This model is widely used in models of distributive politics, and it has been imported from there to models of party formation (Schwartz, 1989; Aldrich, 1995) and ideology (Bawn, 1999). The use of the framework is thus directly comparable to these "long coalition" models, in which stable coalitions emerge as equilibrium strategies in a collective choice framework.

Table 1 provides an example. In it, there are five actors, labeled A through E, who have positions on five issues, numbered I through V.

In this case, $\beta = \{1, 1, 0, 0, 0\}$ is an ideology, because actors A and B both hold it. Technically, $\beta = \{1, 1, 0, 0, 1\}$ is also an ideology, held by only actor C. But more interestingly, actor C *almost* holds the ideology of A and B. If we applied a scaling technique to these data, actor C would not

Table 1: Hypothetical Intelligentsia with Five actors and Five issues

	I	II	III	IV	V
A	1	1	0	0	0
B	1	1	0	0	0
C	1	1	0	0	1
D	0	0	1	1	1
E	0	0	1	1	1

have the same ideal point as actors A and B, but it would be close. In a one-dimensional model, C would be a moderate leaning toward A and B. Ideologies define the poles of such a model.

1.2 Beliefs

There is one last feature of the world that we need to develop before turning to its dynamics. That is the certainty with which the actors hold their positions on the issues. Parallel to the positions held (y_{ij}), we can define beliefs they have about those issue positions, as follows:

- $b_{ij} \in (0, 1)$ is the i^{th} actor's beliefs about the j^{th} issue.

b is constructed to have a correspondence with issue positions, as follows:

$$y_{ij} = \begin{cases} 0 & \text{if} \quad b_{ij} < 0.5 \\ 1 & \text{if} \quad b_{ij} \geq 0.5 \end{cases}$$

Positions are binary, but beliefs are not. Beliefs closer to 1 correspond to $y_{ij} = 1$, while those closer to 0 correspond to $y_{ij} = 0$. Those with beliefs at exactly 0.5 are indifferent, and for the model, the tie will go to supporting the policy. Note that we are using the term "beliefs" in the Bayesian sense as beliefs about the state of the world. We are not referring to the notion of beliefs about the other actors in the game. There is nothing in the model that implies that beliefs has any connection to "the truth." Actors may believe that they are attempting to find the "right" or "just" position, but there is no defined "truth" in the model.

A player's beliefs may be subject to influence and change. For the purposes of this model, we will consider communication among the actors as a way to change beliefs.

Finally, we would expect ideologies to involve a high degree of certainty. At least empirically, survey respondents who are very ideological also hold their issue positions with more confidence than do those who are not ideological.

2 Why Constraint?

If all we are interested in is the implications of ideology in some larger domain, the notion of constraint is sufficient. But we might also want to know why one particular constraint is the one that we observe. And so there are numerous schools of thought on what holds together an ideology. Three stand out. One holds that ideology derives from a psychological or cultural trait, something buried in the way we think about politics. A second holds that ideology is a rationalization or justification for interests, or "shabby motives" (Apter, 1964). And a third holds that ideology is some way of linking actions to principles.

2.1 Traits

A rich literature (e.g., Adorno et al., 1950; McCloskey, 1958; Constantini and Craik, 1980, see Jost et al., 2003, for a more complete review) argues that ideology has psychological roots in personality traits. Some people are more dogmatic, have a strong need for order, have a higher fear of threats and greater moral salience. These people are more likely to be conservative. Others have high tolerance for uncertainty, are more open to new experiences, can manage complexity, and tolerate system instability. These people are more likely to be liberal. Scholars have found strong correlations between physiological measures and ideological self-identification (Kish, 1973; Alford, Funk and Hibbing, 2005; Amodio et al., 2007; Mendes et al., 2007; Oxley et al., 2008). For example, Oxley et al. (2008) find that conservatives have a higher galvanic skin response to threats.

These findings suggest that one of the things that drives ideological differences is just that humans have different predispositions.[1] These predispositions are still filtered through the social and political system, so that actors do not admit to (if they are even aware of) being drive by

[1] A related question is whether such predispositions have genetic roots or are socialized. For the purposes here, all that matters is that these traits precede political behavior.

their biological urges. They instead appeal to general principles and norms, perhaps without fully embracing them.

While the notion of traits underlying ideology has considerable evidence, it also leaves some questions unanswered. For one, these underlying traits appear to be nonlinearly related to a primary ideological dimension.

And in fact, psychological traits do predict a multidimensional ideology. Feldman and Johnston (2009), for example, show that among ordinary Americans, there are two distinct ideological dimensions, an economic and a social dimension. They show that predispositions about equality are related to the economic dimension, but not the social. And predispositions about morality are related to the social but not the economic. Similarly, many of the traits we associate with conservatism are nonlinearly related to a single ideological dimension (Greenberg and Jonas, 2003). Those with the highest levels of aversion to change, for example, may be conservative, but moderates have the lowest levels. The curve bends back up among liberals, but it does not reach the same level as it usually does among conservatives. Moreover, preferences tend to be more multidimensional-dimensional among the less ideological and less sophisticated. That suggests a process in which some aspect of unidimensionality either requires more sophistication, or else the more sophisticated are more likely to be exposed to unidimensional messages.

In the model presented in section 1, these traits might be seen as shaping the initial beliefs (b) of the actors. Some are inclined, on a precognitive level to favor one policy over another. In particular, traits might induce certain correlations among the confidence levels, so that those with certain traits will have similar priors across a number of issues. That could translate into patterns among issue positions that we observe as ideology.

But the pattern would need to pass from the precognitive beliefs to the preferences. It might be overridden, or not activated in certain contexts or cultures. Consider the subjects in the Hibbing et al. study who are more likely to express fear from external threats. These subjects might have a prior for any policy that promises to defend them from some such threat. But on further reflection, or in the absence of a political system that activates that threat, they may change their minds and update their beliefs accordingly.

2.2 Interests

Meanwhile, another tradition (Marx and Engels, 1845; Mannheim, 1955; Downs, 1957; Schwartz, 1989; Bawn, 1999) says that ideology is a rationalization of interests. Policies make winners and losers. Those who will be made less well off by a policy will wish to oppose it. But they may not wish to express their opposition in such naked terms. Instead, they appeal to abstract principles, but apply those principles in service to their own well-being.

As with traits, interests would seem to shape the prior beliefs (b) of the actors in the model above. When an issue presents itself, an actor's first impulse is to favor the policy that would best serve them. But, again as with traits, this belief would need to be translated into a position. Actors can be persuaded that the policy that would best serve them might still not be "just" or best for the "common good." There is considerable evidence (e.g., Sears et al., 1980) that people do favor policies that are against their objective self-interest.

The role of interests becomes more complicated if we consider the potential for logrolls. Bawn (1999) argues that ideology might be seen as an equilibrium in a policy setting game. Actors in such a game might enter into a "long coalition" (see also Schwartz, 1989; Aldrich, 1995) with others, enacting a permanent logroll. The long coalition induces the actors to have preferences on issues they otherwise have no stake in.

One problem with Bawn's argument, however, is that the equilibrium selection mechanism is highly unrealistic. Bawn's model involves a group choosing the coalition, as in a legislative setting. This is an apt model for say selecting a party's platform, but ideologies evolve. Ideologies are also bound together by a more subtle glue. They appeal to principles, not loyalty. At least in theory, such principles could induce actors to choose policies that violate their self-interest or their predispositions.

2.3 Principles

And so we turn to principles. For many, ideology is not just something that scholars study but something that they sometimes do. Ideological thinkers at least aspire to be developing good theories of social order and justice, seeking to persuade others that they are right. Even if scholars believe ideologues are limited by their personalities or their self-interest, there is likely some consequence of their efforts to rise above those limits.

As noted in both the case of interests and of traits, actors appeal to principles and values in defending their positions. In those accounts, principles are essentially window-dressing, concealing the true source of preferences. But why is such window-dressing so appealing? Most people would hold the normative position that something like principles and values *ought* to guide our political positions. If so, it is reasonable to conclude that some people, attempting to abide by this norm, actually succeed.

In the model outlined above, such principles would be unlike priors. Rather, they are something that guides us across many issues. So they would influence the β's in the model above.

We would thus like a way to formalize the process of thinking about principles and values. That is, we would want to formalize the process of normative political theory. We might think this is an absurd task: How do we formalize the creative process? However, we can turn to one of the giants of normative theory for an approach that can at least allow us to model some of the features of the process.

3 Reflective Equilibrium

Throughout his career, John Rawls attempted to develop both a normative theory of justice as well as the reasons why we should accept that theory. We are less interested here in his "principles of justice" or even the "original position" used to justify them. The focus here will be on the broader process called "reflective equilibrium," which we will define shortly. Since ideologues are essentially trying to do political theory, we can look to what political theorists are attempting to do as a model for ideological thinking.

3.1 Who Participates

This requires that we think rather highly of ideologues, putting Rush Limbaugh on the same plane as Robert Nozick. Certainly, most ordinary Americans who are very ideological do not meet the standard of sophisticated political theorists. However, as noted above, we are restricting ourselves to those who are attempting to articulate belief systems. This means we are looking for a model to explain the work of public intellectuals and opinion makers such as Hannah Arendt, Fred Barnes, William F. Buckley, Noam Chomsky, Herbert Croly, Simone de Beauvoir, Milton Friedman, David Horowitz, John Maynard Keyes, Morton Kondracke, Charles Krauthammer, Bill Kristol, Irving Kristol, Paul Krugman, John Leo, Anna Quinlan, William Saletan, Robert Samuelson, Herbert Spencer, Gertrude Stein, William Graham Sumner, Lester Ward and George Will. The model probably will not explain the preferences of people like George Bailey, Willie Loman, or Homer Simpson, because they would take cues from ideological elites but not participate in forming ideologies.

The intellectuals named here are not all equally respected in academic circles. For the purposes of this argument, we must consider them in a single category. Some may be smarter, more introspective, more willing to question themselves or otherwise be better philosophers. I will claim that this variation will cancel out, at least insofar as they are responsible for shaping ideology. But it is necessary to treat them equally if we want to be detached ourselves, since different scholars might differ over who is more "legitimate" as a thinker. Michael Moore and Ann Coulter are probably just outside the lower bound of such a group, but for our purposes, if one is in the intelligentsia, both must be.

3.2 How it Works

The method I propose for their work is the "reflective equilibrium," advanced by Rawls (1971, §9; 2001, I§10; see also Daniels 1979). The model is one adapted from the practice of positive science (Goodman, 1955), in which scientists work back and forth, inductively and then deductively from theory to data, to develop useful theories of the world.

Reflective equilibrium is not an equilibrium concept in game theory. What follows here is not an attempt to develop such a concept and apply it to a game. Rather, this paper develops a game that reflects a thinking process like that Rawls describes, and that has short-run and long-run properties that we can compare to reflective equilibrium.

Rawls argues that humans have a "capacity for reason" and "a sense of justice." Using these, they reach judgments about what is right and wrong. These judgments are not unimpeachable. Some ought to be reconsidered. But some, which Rawls calls "considered judgments or considered convictions," are so well founded that they ought not be questioned lightly. "Some judgments we view as fixed points: ones we never expect to withdraw, as when Lincoln says: if slavery is not wrong, nothing is wrong " Rawls (2001, p. 29).

Even so, some of our judgments are in conflict with one another. The philosopher resolves these conflicts by finding "the conception of political justice that makes the fewest revisions" in those initial judgments. This resolution, in which "general convictions, first principles, and particular judgments are in line" is a "narrow reflective equilibrium" (p. 30).

Consider a brief illustration: Following Rawls' example, Lincoln offers slavery as an issue on which we should have a considered judgment. The South's holding of slaves was unjust, and warranted policies to eventually end slavery.[2] Based in part on this opinion about slavery, some thinkers in the late nineteenth century reasoned that Spanish domination of Filipinos and Cubans was wrong. The reasoning moves from the specific, which helps to inform the general, and then apply to another specific.

More than this narrow reflective equilibrium, however, Rawls says we should pursue a "wide" reflective equilibrium, which is "reached when someone has carefully considered alternative conceptions of justice and the force of various arguments for them" (p. 31). Finally, Rawls argues that since everyone is working toward a public conception of justice, we all should arrive at the same wide reflective equilibrium.

The distinction between narrow and wide reflective equilibrium is where Rawls introduces something that might distinguish the casual thinker from the true philosopher. *Real* political theorists attempt to engage in wide reflective equilibrium. And such an approach might allow them to transcend their myopic interest-based and psychological limitations. This paper, however, considers the work of those who, at least in the short term, are still working through narrow reflective equilibrium. Even if we believe that Rawls is right, and that wide reflective equilibrium will eventually lead us to a unique (and correct) theory of justice, it is empirically the case that in 2010, we have

[2] Of course, that southerners did not see the issue that way calls into question the use of the issue as a reference, a point to which we will return below.

not yet arrived. Today, everyone is at best only just moving from narrow to wide, and many are still working on the narrow. And reasonable people attempting to arrive at theories of justice will, in the short run, differ on those theories. It is exactly this limitation that I think explains different ideologies.

3.3 Competing "Considered Judgments"

The key is in the need to decide which judgments are "considered" and which are merely strongly held. In a political system, such disagreement is common. Worse, most of the time, thinkers will not agree on when we do not have a considered judgment. For Rawls, you should only use those judgments that are widely shared. But who decides when reasonable people agree?

Consider the application of the same model in the hard sciences. As a model of science, reflective equilibrium is on solid ground. Historically, scientists debated over a heliocentric versus geocenteric model of the universe. Both are based on a theory that says the earth, the sun, and other planets are heavenly bodies moving through space. And both understand that the earth and the sun revolve around each other. But which is the real center? Which do other bodies like the moon and planets revolve around? Enter retrograde motion, the observed short-term reversal in the paths of planets through the sky.

The heliocentric model is consistent with retrograde motion, while the geocentric model is not. So we reject the geocentric model and are forced to develop the heliocentric. That is, we keep it until we observe that other objects do not seem to be revolving around the sun. And we discover that while the sun is the center of the Solar System, it is not the center of the galaxy, or the universe. And the scientific refinement of theory continues.

But consider a parallel case in philosophy. Americans embrace "freedom" as a principle. But freedom might have different forms. Negative freedom (or freedom as self-ownership) says that we want to minimize the demands the state makes on individuals. Positive freedom (or freedom as optionality) says we want to maximize the opportunities that individuals have to achieve their goals. Since these models are often in conflict, which is right? Enter labor laws. Labor regulations are not consistent with negative freedom but are with positive freedom. However, labor laws are not like retrograde motion. While we all must agree with what we see through the telescope, judgments about the value of labor regulations are not universal.

For Rawls, this need not be a problem. Labor is just not a case on which we have our "best considered judgments," so we will not use it to guide our choice. But for many, labor is a solid judgment. A small business owner, for instance, knows all too well the consequences of giving too much power to workers, and if anything is wrong, it is denying a business owner the ability to run her own business her own way. Workers are always free to leave. Likewise, a factory worker may have few judgments he considers more settled than that he should never be put in a position where he has little option but to take a job that does not treat him safely and compensate him adequately. For him, if anything is wrong, undermining his ability to unite with his coworkers is wrong.

"What gets us into trouble is not what we don't know," observes Mark Twain. "It's what we know for sure that just ain't so." Most political issues are like this. Indeed, Lincoln's slavery may have been a considered judgment to him, but it was not to half the country. It is not simply that the issues are unsettled. It is that some people think they are settled, and different people think they are settled on different sides. This is because different people have different interests and psychological predispositions. Even something as simple as different experiences can lead to different judgments. This is true even for the most sophisticated thinkers, who are trying, contentiously, to find the right philosophy.

In fact, Rawls himself argues that we should expect people to have competing values and beliefs because of the "burdens of judgment." Specifically, the evidence we weigh is complex, we can disagree about the weight placed upon different considerations, political concepts can be vague and subject to hard cases, our reasoning is shaped by our experiences, and different kinds of considerations are hard to compare. Rawls does not link the concept of the burdens of judgment to the concept of reflective equilibrium in the way that I have, but the relationship is there. For this reason, it may be that we do not all agree. Rawls wants to focus on how we might eventually come to agree. I want to focus on what happens in the meantime, when we do disagree.

There is some evidence that political thinkers do follow this process, and that they can be especially attached to judgments, even as the principles that justify them change. Consider a few examples.

In his own telling, G.A. Cohen, a Marxist social philosopher, had long agreed with many Marxists that the cornerstone of capitalist exploitation was that the capitalist extracted something from the worker that the worker himself owned, his labor, and a person cannot be coerced into giving up something that they own. Freedom was thus about self-ownership. However, in arguing with those who supported capitalist economics (notably Robert Nozick), Cohen came to believe that the principle of self-ownership did support a minimal state and not interventions on behalf of equality. Cohen thus rejected self-ownership (the principle), since it doesnt justify the egalitarian policies Cohen has firmly judged to be right. (Cohen, 1995).

Rawls himself is tied to some very specific judgments. In 1958, in "Justice as Fairness," Rawls used a variation on game theory with rationally expectant actors to justify his two principles[3] of justice. In 1971, in *A Theory of Justice*, Rawls had decided that the two principles were not implied by standard rational-actor game theory, and instead developed his well-known idea of the "original position" behind the "veil of ignorance." In 1992, Rawls took a more intensely pluralistic approach to political theory, but again arrived at the same two principles in *Political Liberalism*. The same principles survived with minimal revisions into *Justice as Fairness: A Restatement* in 2001.

In both of these cases, it is not that the theorists are being insincere. They are genuinely attempting to reach a conception of what is right. The judgments they hold fast to are not blind spots, but the things they believe most strongly to be right. But reasonable people can disagree about what those things are. Or at the very least, empirically, actual people who claim to be reasonable do disagree.

Thus the model of reflective equilibrium may not work to reach the correct answer, at least in the near future. The problem is, psychological predispositions and interests color our judgment on which issues are settled. Why introduce reflective equilibrium, then? I argue that ideologues are trying to reach what they think is a correct answer, and that they are applying something like reflective equilibrium. Even when they think they are not, they often are, in the same way that

[3] Briefly stated: 1. All persons have a claim to equal liberty, and 2. Any social and economic inequalities are justified only if all offices and positions are open to all and only if inequalities are to the advantage of the least advantaged members of society.

political scientists who claim to be deducing theories from axioms are still influenced by empirical observation.

But using a flawed system has consequences. To understand them, we need a model that puts structure on the limitations from interests and psychology.

4 The Game of Reflective Equilibrium

This section lays out the basic game play. Details and proofs are in the appendix.

The game involves a group of players called the **intelligentsia**, with preferences and beliefs as described in section 1. The players are all engaged in the process of working out positions on all of the issues of the day. As noted above, this is a model of the "miniscule proportion of the population" that engages in creative synthesis. These actors are interested in arriving at positions on issues that they regard as "just," "right" or "true." They have their own beliefs on the issues, but they may communicate with others and therefore update those beliefs.

Before we get to the game, consider a possibly ideal scenario, in which all actors get together and discuss their beliefs, perhaps voting on every issue. We know from the Condorcet Jury Theorem (Condorcet, (1785) that, *if the actors' beliefs are reflective of the truth, and if the sources of their uncertainty are independent*, such a procedure would be very effective. That is, if the probability that each actor's position on the issue is "correct" is greater than 0.5, then the probability that their aggregate decision from majority rule is also "correct" goes to 1 as the number of actors grows large.

But the actors in this scenario cannot appeal to the Condorcet Jury Theorem, for two reasons. First, there is no objective truth in this model. So their positions cannot have any relationship to it. Second, the actors are trying to develop a philosophy about the world that is more than just majority rule. Perhaps they recognize that, on some issues, the prevailing wisdom does not reflect the truth. At any rate, their goal is to individually develop a set of principles that point to justice. They will work with others, but in a prescribed way.

Their interaction with each other mirrors the way in which ideas are published. The actors write down their theories, then attempt to share them with the world. To do so, they face a gatekeeper, drawn from the rest of the intelligentsia. The process repeats.

4.1 The Stage Game

The stage game is modeled on the peer review process. A political thinker, called the **Theorist**, proposes a theory, which has the form of an ideology. That is, it is a vector of positions on all issues. Another thinker, the **Reviewer**, considers, then either accepts or rejects it. The "peer review" process a useful metaphor, but only a metaphor. In broad strokes, it captures the way in which political discourse is conducted. For example, in the case of political journalists (writing in say *The New Republic* or *The Weekly Standard*), we would use the terms "Contributor" and "Editor." Sometimes in political journalism, the decision to accept or reject comes at the hiring stage, with the decision to give someone a column. But the main idea remains that the speaker and the gatekeeper are from the same class, often working on both sides of the fence, and serving as both actor and (part of the) audience.

The game has two steps. First, a Theorist is chosen from among the polity to propose a β. This is a complete ideology, with prescriptions for all issues. It is thus a vector of 0's and 1's. Second, a Reviewer is chosen from among the polity. That Reviewer gets to "accept" or "reject" the piece. After its acceptance or rejection, both actors update their beliefs about all issues.

Initially, we will consider the case where the Theorist and Reviewer are chosen randomly and independently. I will then consider the more realistic case in which the Reviewer's preferences are likely to be correlated with the Theorist's, on the grounds that the Theorist chooses her outlet that has some similarity to her own preferences.

The model also does not consider what happens after acceptance or rejection, except as concerns the Theorist and the Reviewer. Other thinkers do not read what is published except when they are acting as gatekeepers. In reality there are no doubt other thinkers who read, but they are similarly skeptical, in the same way as a reviewer, and only believe what they would themselves accept.[4] Any further audience for political journalism occurs at a separate stage, involving readers who are consumers but not participants.

[4] Consider by way of another metaphor the RAS model of Zaller 1992, but as conducted by highly informed actors who are thus resistant to inconsistent messages.

I assume the actors play the stage game strategically, but without regard to its long-term considerations. Utility is assigned at the end of the stage, and except for the evolution of beliefs, previous rounds outcomes do not affect the utility at future rounds. Actors are not strategic about the long-term evolution of beliefs.

4.2 Prior beliefs and Preferences

Actors have lexical preferences over issue areas: They care first about issues which they regard as considered judgments, and only then about other issues. This captures a core notion of the role of considered judgments in Reflective Equilibrium: that we should view our best considered judgments as "fixed points,"[5] which we expect never to withdraw.

We define a **considered judgment** as any issue on which an actor has reached some level of certainty. Using the conventional level of confidence in hypothesis testing, that would be any issue on which an actor was 95 percent certain. The model here uses that level. When beliefs cross that threshold, the actors do not reconsider their beliefs. It is as if they are 0 or 1, and the model treats them that way.

The preferences of the actors over the outcomes of accept and reject are detailed in appendix A.3. The preferences involve two important features:

First, the actors care about both substance of what is accepted or rejected, and whether or not something is accepted or rejected. That is, there is a premium for publication, but the actors also care about what they say (or allow to be said). This creates a tradeoff, in which a Theorist may say something she does not necessarily believe in order to get it accepted.

Second, this tradeoff does not apply to considered judgments.

Appendix A.3 outlines the microfoundations consistent with these preferences, but if the reader is skeptical that actors might be able to work through the tradeoffs as described, all that is important is that these features give the following equilibrium strategies:

[5] Here "fixed point" is used in the way Rawls used it, as a guiding reference point, and not in the way mathematicians usually use it, to refer to a point that is mapped to itself by a function.

The Theorist presents a theory that matches the Theorist's preferences on all considered judgments, but may bend to popular will on other issues, depending on the tradeoff.

The Reviewer accepts any theory that matches the Reviewer's preferences on all considered judgments.

4.3 Updating beliefs

After acceptance or rejection, both the Reviewer and the Theorist have learned something.

For the Theorist, if the Reviewer accepts, the Theorist should become more confident. The Theorist knows that the Reviewer doesn't necessarily agree with everything presented, but on balance, everything is more likely. If the Reviewer rejects, the Theorist becomes less confident (except for those issues which the Theorist believes are settled —her considered judgments).

For the Reviewer, something similar happens. If the Theorist offers something that agrees with him on all considered judgements, then the theory is persuasive in general. Thus the Reviewer's beliefs move toward the theory on all issues. However, if the Theorist offers something that is disagreeable, then the Theorist has proven herself untrustworthy, and the Reviewer's beliefs move away from the theory.

Just how fast the beliefs move in these directions is another matter. It would be nice to work out a result following Bayes Rule. However, there is no objective moral reality in this model. To compute the posterior probability for an issue position given that it was proposed by the Theorist, for example, we would need to know the objective probability that the Theorist would propose it given that it is true. If no one knows the true moral world, however, then no one can know the relationship between the Truth and anyone's beliefs. Appendix A.6 spells out a solution, in which beliefs move in the suggested direction, but not too quickly.

5 Simulation Results

This section presents results from three versions of the model. I present them in parallel for comparison. The first version of the model is a baseline, as described above.

The first variant modifies the initial beliefs to have some structure. This echoes the kind of structure that psychological theories suggest may underly ideology. Actors have some tendency toward one of three preference profiles, say perhaps an initial bias toward policies that favor a strong state, or that punish wrongdoing.

The second variant leaves beliefs as they are in the base model and instead modifies the selection process for Reviewers, so that they are disproportionately likely to agree with the Theorist. If we think, for example, that Theorists seek out favorable outlets for their views, then they will not have to suffer disagreeable Reviewers

Each version of the model has been simulated 50[6] times for 500 iterations of the stage game. In the simulations, there are 20 actors and 20 issues. Code for implementing the stage game of the model in R is provided in Appendix B. Code for generating and storing many iterations of the model and generating the figures here is available from the author.

There are several features of the simulation that we would like to consider.

5.1 Certainty

Reflective Equilibrium is meant to form principles that will give confidence to our convictions. The way this model is laid out, certainty is an absorbing property.

Figure 1 traces the number of issues held with certainty over the course of 500 iterations. The mean number is plotted, along with a 95 percent frequency interval. The upper left panel shows the base model. To the right of the base model is Variant 1, in which the initial preferences were given some structure. Below the base model is Variant 2, in which Reviewers are likely to have similar preferences to the Theorist. All other figures are laid out in this way.

Across all three models, the percentage of positions held with certainty grows. It grows the fastest when Theorists get agreeable Reviewers. Such a system is self-confirming. Clusters of like-minded thinkers talk to one another and reaffirm their beliefs, never having to question themselves.

[6] Larger simulations are planned.

5.2 Disagreement

The process of Reflective Equilibrium should bring about agreement across actors. But, as I have argued, different considered judgements should keep actors from reaching agreement. Still, they should come to agree with some others. Empirically, we observe political conflict among a small number of positions with many adherents.

Figure 2 shows the distribution of unique preference profiles held by actors at the end of 500 iterations. That is, the 0's and 1's held at the end of the simulation. What we find is that, for the base model and Variant 2, the total number of unique profiles has fallen. And even for Variant 1, there are many runs with very few unique profiles. At the same time, it is not that common to have only one profile at the end of the system.

5.3 Dimensionality

Figure 2 suggested that there are a small number of unique preference profiles, but that distinguishes between profiles that are very similar. There may be many moderates, but still two dominant preferences —or at least two poles. In modern politics, we generally observe two ideologies in conflict, defining a single-dimensional ideological space. This in spite of the fact that many issues need not be connected.

There are many ways to measure dimensionality. I apply three here, all using the eigenvalues from the matrix of preferences. Figure 3 traces the mean value of the first eigenvalue. Large first eigenvalues suggest that one dimension significantly organizes the preferences. And figure 4 plots the average ratio of the first and second eigenvalues. As this number gets large, one dimension is dominating over the others.

Again, across the variants, the dimensionality grows over time. In the case of the agreeable reviewers, the ratio of the first to second eigenvalue explodes after a handful of iterations. (Note the scale on the lower left figure.)

The average is somewhat misleading, however, because it smooths out variation over time. The process is actually more volatile, as we can see from Figure 5, which plots the ratio of the first two eigenvalues for a single run of 500 iterations. Here, we see that dimensionality waxes and wanes, presumably as different issues become associated with each other and then not.

Why does this occur? It's difficult to tease out the exact contribution of every facet of the model, but it is reasonable to see that subsets of the intelligentsia move toward each other and away from other subsets. As more of their positions become fixed, they are unable to move back toward others. If we imagine them having different ideal points in a multi-dimensional space, to the extent that this smaller number of subsets are aligned, they will fit on a single line. If two poles are frequent, we should expect a strong single dimension.

5.4 Polarization

Finally, not only do we observe two dimensions, we also observe polarization. We can measure polarization by estimating a scaling technique, such as NOMINATE at each iteration. We can then take the ratio of the interquartile range to the total range of the first dimension NOMINATE score. Where this is high or close to 1, we have polarization. Figure 6 plots this over time, and figure 7 plots it for a single iteration. We do not see a general tendency toward more polarization, but there is considerable variation in polarization.

6 Discussion

The results from simulations of this model seem to match the stylized facts of the real world. Reflective Equilibrium under these conditions tends toward a small number of ideologies, often on one dimension. This is not a test of the model so much as a reality check that what is being modeled is captured. Further comparative statics would be useful.

6.1 Extensions

The number of potential extensions to this model is large. Here are a few that would be worth exploring.

6.1.1 Awareness of the total coalition

Part of the incentive to form a long coalition comes from the awareness of its members that there is some threshold for victory. Nothing like that is included in the current model. But it could be. Alternative decision rules for the theorist, for instance, might account for how many people might be persuaded by the proposal. Theorists could condition on the preferences of those who agree on considered judgments, for instance. One way to move in this direction might be to have proposed theories be voted on by the entire intelligentsia, rather than just one gatekeeper.

It would be useful to derive decision rules assuming utility functions over the number of people who agree with the proposal. Theorists presumably prefer to propose their own beliefs, but they might also derive utility from advancing popular theories. In that case, when their own beliefs are not strong, taking cues from the others makes sense, and proposing theories that are more likely to lead to long coalitions will be more likely

6.1.2 Broad dissemination

Relatedly, all players should be more aware of the other ideologies that are being offered. One natural extension is to let everyone update on the basis of any β that passes peer review. Likewise, peer review could be based on more than one peer. The more peers required, the more likely a proposal will be rejected, and thus the greater the updating if it is not rejected.

6.1.3 Reconsidering considered judgments

Under the current model, once a thinker reaches some threshold of certainty, they have a considered judgment and they never reconsider. It would be easy, however, to allow reviewers to act as though they were certain once they reach such a threshold, but still allow their certainty to cross back over the threshold, so as to be persuadable in the future. This would be more consistent with the notion as Rawls lays it out.

6.1.4 Salience

Under the current model, all issues are equally salient. But one reason long coalitions can form is that different actors care about different issues. Degrees of belief captures this idea in the current model, but it might be possible to allow members to reach a considered judgment on an issue that they nevertheless feel is less important than some other considered judgment.

6.1.5 Tracing Ideologies vs. Preferences

The current model traces changes in Preferences. But the actors are proposing ideologies that might differ from their own preferences. It is these ideologies that would be disseminated to the public, so it would be worthwhile to trace patterns in them —the β's —as well as with the preferences, and compare them.

6.1.6

A Appendix

A.1 Players

The players are an **intelligentsia**, as described in section 1:

- X is a set of k binary issues.
- P is a set of n intellectual actors.
- $y_{ij} \in \{0,1\}$ is the i^{th} actor's "position" on the j^{th} issue.
- β is a **belief system**, or a set of issue positions that are meant to be internally consistent.
- $\mathbf{Y}$ is an n by k matrix of issue positions.
- $b_{ij} \in (0,1)$ is the i^{th} actor's beliefs about the j^{th} issue.

A.2 Game Sequence

The game proceeds as follows:

1. Nature chooses the beliefs of all of the actors.
    - In the base model, this choice is random from the uniform distribution.
    - In the variant 1 model, this choice is random from a normal distribution centered on 0.6 or 0.4, to create one of three profiles. This models genetic tendencies.
2. Nature chooses which issues are considered judgments for all actors.
    - In the current models, the i^{th} actor treats the i^{th} issue as a considered judgement.
    - In future variants, patterns in the considered judgments will be explored.
3. Nature chooses one actor to be the **Theorist** and one to be the **Reviewer**
    - In the base model, each actor has equal probability of being chosen at each stage.
    - In the variant 2 model, the actor is chosen at random, and the probability of being chosen as a reviewer is a function of how much each potential reviewer agrees with the chosen theorist.
4. The Theorist presents the theory
5. The Reviewer accepts or rejects
6. Both the Theorist and the Reviewer update their beliefs.
7. Both the Theorist and the Reviewer form new considered judgments for any beliefs that have just crossed the threshold.
8. The game repeats steps 3 through 7 indefinitely.

A.3 Preferences

A.3.1 The Theorist

Begin with the preferences for the Theorist. On the first question, considered judgments, the Theorist prefers to say what she believes strongly to be true over saying what she believes strongly to be false. This implies the following:[7]

$$U_T(\beta_i = b_{T,i}) = 1 \tag{1}$$

$$U_T(\beta_i \neq b_{T,i}) = 0 \tag{2}$$

Because the Theorist's preferences are lexical and with this form, we can pause here to note that no Theorist will ever propose a β that contradicts their own considered judgments. This is an assumption with some bite. It rules out two possibilities. First, it rules out the purely opportunistic theorist who will write what he "knows" to be false just for attention. If publication itself is highly valued, such a tradeoff might make sense. Second, it rules out the slightly less opportunistic theorist who might trade off saying one thing she "knows" to be false for the opportunity to say many other things she "knows" to be true.

While there are surely such opportunistic actors (some might say evidenced by most political best-sellers), in the domain of serious theorists, whose work is likely to be taken most seriously, it is plausible to make this claim. The goal of this paper is to take political thinkers at their word, and that means assuming they are at least sincere when they have strong beliefs. Note that the claim applies only to those issues which the actor views as a considered judgment. She may well sell out on any other.

After weighing considered judgments (and so not violating her deepest held convictions), the Theorist's preferences are over what she says and whether it gets published (whether it is Accepted $= A$):

[7] Since the beliefs of an actor on those issues she considers a considered judgment are a 0 or 1, and since the β is always a 0 or 1, most loss functions based on distance (e.g. quadratic or linear) would reduce to these expressions.

$$U_T(\beta|A) = -||\beta - b_T|| + \theta(\gamma - ||\beta - b_T||) \tag{3}$$

$$U_T(\beta|\neg A) = -||\beta - b_T|| \tag{4}$$

where θ is a scalar that captures the added value of having ones opinions published, and γ is a scalar that captures the pure value of publication regardless of the content. One might imagine θ varying by outlet and γ varying by actor, although those possibilities are not explored here.

A.3.2 The Reviewer

The Reviewer's job is much easier. The Reviewer also has lexical preferences that weigh his considered judgments first. He thus also first prefers to accept those arguments that match his considered judgments and reject those that do not.

This is the only consideration of the reviewer. He otherwise prefers to accept all articles. He thus accepts anything that does not violate a considered judgment. This makes the reviewer especially open-minded, perhaps best modeling a typical audience member for the piece. If the piece passes muster on the issues on which the Reviewer is passionate, the Reviewer is highly receptive to the rest of the argument.

This open-mindedness is also in keeping with the notion of Reflective Equilibrium. Actors should expose themselves to a wide range of ideas.

A.4 Knowledge

The players do not know each other's preferences, but they do know aggregate preferences on each issue. (We might imagine regular polling of elites, or an unbiased sense of issue preferences from interaction at cocktail parties.) That is, they know the percentage of all pundits whose beliefs are greater than 0.5 for each issue. Call this vector a. The Theorist does not know how correlated the preferences on these issues are. That is, she does not know probability of support for an issue conditional on support for another issue. But she does know the marginal probabilities.

A.5 Stage game equilibrium strategies

I solve by backwards induction. At the second stage, the Reviewer accepts any β that does not violate his considered judgments. The Theorist prefers the be published than not (conditional on what is being published) and would thus prefer to offer a β with no contradictions,. The Theorist does not know the preferences of the Reviewer, but she can compute the probability that any given β will be objected to. Since that probability is just a function of β, the Theorist's problem is to choose β to maximize the following expression.

$$U_T(\beta) = U_T(\beta|A)p + U_T(\beta|\neg A)(1-p) \tag{5}$$

where the probability that the Reviewer agrees to publish is the probability that there is no disagreement with a considered judgment, which is given in equation 6, where q is the base probability that any issue is a considered judgment, which the Theorist can infer from her own rate of certainty across many issues.

$$p(\beta) = \prod_i^n 1 - q\Big(1 - \sqrt{(\beta_i - a_i)^2}\Big) \tag{6}$$

Equation 6 could be maximized by matching β to the population's beliefs a. That is, the Theorist maximizes publication by saying what she thinks most people want to hear. Of course, the Theorist does not merely want to maximize publication. She also cares about what she says, which is captured by Eqations 3 and 4 above.

Combining equations 3,4 and 6 via 5 yields 7:

$$\theta\gamma\Bigg[\prod_i^n 1 - q\Big(1 - \sqrt{(\beta_i - a_i)^2}\Big)\Bigg] + \theta\Bigg[\prod_i^n 1 - q\Big(1 - \sqrt{(\beta_i - a_i)^2}\Big)\Bigg]\big(-||\beta - b||\big) + \big(-||\beta - b||\big) \tag{7}$$

which the Theorist then maximizes, subject to the constraint that β does not disagree with any of the Theorist's considered judgments.

Equation 7 cannot be easily solved analytically, but it can be solved computationally. Even without analytic comparative statics, we can say something about the properties of this solution.

As γ gets large, the relative importance of improving the probability of acceptance grows. So the more the Theorist values publication, the more she will compromise her own beliefs. However, θ affects both the first and second term, so the quality of the journal affects both the importance of compromise but also the importance of what the Theorist says. Both of these are as we would expect.

A.6 Updating beliefs

After the game, both actors update their beliefs, based on the information revealed in the game. For the Theorist, if the piece is accepted, that should confirm her beliefs in the things she has said, whereas if it is rejected, it should reduce her beliefs. For the Reviewer, if the piece is accepted, he updates on the new things that were not the basis of the acceptance (were not considered judgments). If he rejects the piece, the Reviewer's beliefs move away from the proposed β.

Ideally, updating would follow Bayes Rule. However, because there is no objective moral reality in this model, Bayes Rule cannot be applied. To compute the posterior probability for an issue position given that it was proposed by the Theorist, for example, we would need to know the objective probability that the Theorist would propose it given that it is true. But we have assumed no relationship between the Truth and beliefs on these thorny moral issues. So we can't define the objective probability in equation 8. That is the nature of the political philosophy problem. That is, there is no way to define $p(\beta_i = 1|y_i = 1)$ and similar terms in the following expression of Bayes Rule. The actors are not conditioning on the truth, since they do not know it.

$$p(y_i = 1|\beta_i = 1) = \frac{p(\beta_i = 1|y_i = 1)p(y_i = 1)}{p(\beta_i = 1|y_i = 1)p(y_i = 1) + p(\beta_i = 1|y_i \neq 1)p(y_i \neq 1)} \tag{8}$$

However, we do know something about the plausible relative size of those objective probabilities. In the model, actors are more likely to choose a β and to accept a β if they strongly believe its elements to be true than if they strongly believe them to be false. It is a plausible step for the actors to assume β's are more likely to be proposed or accepted if they are true than if they are not. This assumes a certain minimal belief in the discipline, but one that it is reasonable to assume most theorists do have. The magnitude of the difference cannot be determined, but that merely affects how fast beliefs will change. We can thus argue that actors assume

$$p(\beta_i = 1|y_i = 1) > p(\beta_i = 1|y_i \neq 1) \tag{9}$$

In other words, after the outcome of the game, beliefs move toward the β if it is accepted and away from it if it is rejected. In the model, I assume that the difference is small, so the movement is small (The results reported here assume .45 and .55). It otherwise follows the general form of Bayesian updating.

A.7 Considered Judgments

At the end of the stage game, the actors may have formed new considered judgments. If their beliefs about any issue crosses some threshold, they become certain and do not revisit the issue. The threshold, α, is set to 0.95 in the simulations here.

$$b_t = \begin{cases} 1 & \text{if} \quad b_t > \alpha \\ b_t & \text{if} \quad 0.05 \leq b_t \leq 0.95 \\ 0 & \text{if} \quad b_t < 1 - \alpha \end{cases} \tag{10}$$

B Model code

B.1 Code for Stage Game

```
# PARAMETERS OF MODEL

n <- 20             # NUMBER OF ISSUES
m <- 20             # NUMBER OF PEOPLE
p <- 1              # NUMBER OF CERTAIN ISSUES PER PERSON
alpha <- .05         # THRESHOLD FOR CERTAINTY
T <- matrix(50,n,m)

# DEFINE FUNCTIONS
product <- function(argument) exp(sum(log(argument),na.rm=1))
```

```
bayesrule <- function(prior,A,notA) (A*prior)/(A*prior+notA*(1-prior))

# PATTERNED BELIEFS (for VARIANT 1)
two <- rep(c(.6,.4),length.out=n)
three <- rep(c(.6,.4,.6),length.out=n)
four <- rep(c(.6,.6,.4,.4),length.out=n)
pattern <- append(two,three)
pattern <- append(pattern,four)
beliefs <- matrix(rnorm(n*m,pattern,.001),n,m,byrow=FALSE) # RANDOM, CENTERED ON PATTERN

# INITIAL BELIEFS (for BASE MODEL)
beliefs <- matrix(rnorm(n*m,.5,.01),n,m) # RANDOM, CENTERED ON INDIFFERENT

# FOR SOME RANDOM SAMPLE, CERTAINTY
# MATRIX MEASURING CERTAINTY (1 IF CERTAIN, 0 IF NOT)
certainty <- beliefs*0
for(i in 1:m){  certainty[i,i] <- 1 }  # CERTAINTY DOWN THE DIAGONAL

# INITIAL PREFERENCES (POLICIES FAVORED FROM BELIEFS)
preferences <- beliefs
preferences[beliefs>=.5] <- 1
preferences[beliefs<.5] <- 0

## choose random actors

for (iter in 1:iterations){
#while (qr(certainty)$rank>1){

## SELECT SPEAKER AND PEER REVIEW (for BASELINE)
choose <- sample(1:m,2)

# TO CHOOSE ONE WHO IS MORE AGREEABLE (for VARIANT 2)
Ag.1 <- matrix(0,m,m)
for(eye in 1:m){for (jay in 1:m){
    Ag.1[eye,jay] <- sum(preferences[,eye]==preferences[,jay])
}}
agreeable <- Ag.1[,choose[1]]+.01
agreeable[choose[1]] <- 0
choose[2] <- sample(1:m,1,prob=agreeable)

# THEORIST’S AND REVIEWER’S BELIEFS, DRAWN FROM MATRIX (FOR EASE OF CODING)
t <- beliefs[,choose[1]]
r <- beliefs[,choose[2]]

# Beta is the theorist’s PROPOSAL
# THEORIST'S PROPOSAL = THEORIST'S PREFERENCES (NOT IMPLEMENTED)
# Beta <- preferences[,choose[1]]

# CHOOSE BETA USING PROB. OF ACCEPTANCE
source("chooseBeta_model.txt") # SEE BELOW

bias <- preferences[,choose[2]] # REVIEWER'S BIASES = REVIEWER'S PREFERENCES
bias.certainty <- certainty[,choose[2]] # WHICH ISSUES IS REVIEWER CERTAIN OF?

# DEFINE CONDITIONAL PROBABILTIES: A = PR(b|a), B = PR(b|nota)

## CONDITION: DOES THE PEER AGREE WITH THE PROPOSER'S BETA
accept <- sum(Beta*bias.certainty==bias*bias.certainty)==n
upper <- .55
lower <- .45

if(accept==1){
        ## UPDATE FOR THEORIST IF ACCEPT
t.1 <- (Beta)*bayesrule(t,upper,lower)+(1-Beta)*bayesrule(t,lower,upper)

        ## UPDATE FOR REVIEWER IF ACCEPT
r.1 <- (Beta)*bayesrule(r,upper,lower)+(1-Beta)*bayesrule(r,lower,upper)

## IF REJECT
}else{
        ## UPDATE FOR THEORIST IF REJECT
t.1 <- (Beta)*bayesrule(t,lower,upper)+(1-Beta)*bayesrule(t,upper,lower)
        ## UPDATE FOR REVIEWER IF REJECT
r.1 <- (Beta)*bayesrule(r,lower,upper)+(1-Beta)*bayesrule(r,upper,lower)
   }

t.1[t==0] <- 0
t.1[t==1] <- 1
r.1[r==0] <- 0
r.1[r==1] <- 1

# PLACE UPDATED BELIEFS BACK INTO THE MATRIX
beliefs[,choose[1]] <- t.1
beliefs[,choose[2]] <- r.1

# UPDATE CERTAINTY AND PREFERENCES
certainty[beliefs<alpha] <- 1
certainty[beliefs>(1-alpha)] <- 1
preferences[beliefs>=.5] <- 1
preferences[beliefs<.5] <- 0
```

B.2 Code for choosing beta (called above)

```
Beta.test <- vector("numeric",n)
All <- matrix(0,n,2^n)
for(i in 1:n){
long <- vector("numeric",2^(i-1))
All[i,] <- append(long,long+1)
}

a <- apply(preferences,1,mean)
b <- t
q <- mean(certainty[,choose[2]])
Alpha <- 1
Gamma <- 1

hold <- vector("numeric",2^n)
for(i in 1:2^n){
Beta.test <- All[,i]
p <- prod(1- q*(1-abs(Beta.test-a)))
norm <- -1*sqrt(sum((Beta.test-b)^2))
arg <- p*Alpha*Gamma+p*Alpha*norm+norm
hold[i] <- arg
}

Beta.hat <- All[,hold==max(hold)]
Beta <- Beta.hat
```

References

Adorno, T.W., Else Frenkel-Brunswik, D.J. Levinson and R.N. Sanford. 1950. *The Authoritarian Personality.* New York: Harper.

Aldrich, John. 1995. *Why Parties? The Origin and Transformation of Political Parties in America.* Chicago: University of Chicago Press.

Alford, John R., Carolyn L. Funk and John R. Hibbing. 2005. "Are Political Orientations Genetically Transmitted?" *American Political Science Review* 99(2):153–168.

Amodio, David M., John T. Jost, Sarah L. Master and Cindy M. Yee. 2007. "Neurocognitive correlates of liberalism and conservatism." *Nature Neuroscience* .

Apter, David E. 1964. Introduction: Ideology and Discontent. In *Ideology and Discontent*, ed. David E. Apter. New York: The Free Press.

Bawn, Kathleen. 1999. "Constructing 'Us': Ideology, Coalition Politics, and False Consciousness." *American Journal of Political Science* 43(2):303–334.

Black, Duncan. 1958. *The Theory of Committees and Elections.* New York: Cambridge University Press.

Cohen, G.A. 1995. *Self-Ownership, Freedom and Equality.* Cambridge: Cambridge University Press.

Condorcet, Marquis de. (1785) 1994. *Essai sur l'application de l'analyse a la probabilite des decisions redues a la pluralite des voix.* Paris: .

Constantini, Edmond and Kenneth H. Craik. 1980. "Personality and Politicians: California Party Leaders, 1960-1976." *Journal of Personality and Social Psychology* 38:641.

Converse, Philip. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David Apter. New York: Free Press pp. 206–261.

Daniels, Norman. 1979. "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *Journal of Philosophy* .

Downs, Anthony. 1957. *An Economic Theory of Democracy.* New York: HarperCollins Publishers, Inc.

Enelow, James and Melvin Hinich. 1989. The Theory of Predictive Mappings. In *Advances in the Spatial Theory of Voting*, ed. James Enelow and Melvin Hinich. New York: Cambridge University Press pp. 167–178.

Feldman, Stanley and Christopher D. Johnston. 2009. Understanding Political Ideology: The Necessity of a Multi-Dimensional Conceptualization. In *the annual meeting of the American Political Science Association.* Toronto: .

Gerring, John. 1997. "Ideology: A Definitional Analysis." *Political Research Quarterly* 50(4):957–994.

Goodman, Nelson. 1955. *Fact, Fiction and Forecast.* Cambridge: Harvard University Press.

Greenberg, Jeff and Eva Jonas. 2003. "Psychological Motives and Political Orientation – The Left, the Right and the Rigid: Comment on Jost et al. (2003)." *Psychological Bulletin* 129(3):376–382.

Hinich, Melvin J. and Michael C. Munger. 1994. *Ideology and the Theory of Political Choice.* Ann Arbor: University of Michigan Press.

Jost, John T., Jack Glaser, Arie W. Kruglanski and Frank J. Sulloway. 2003. "Political Conservatism as Motivated Social Cognition." *Psychological Bulletin* 129(3):339–375.

Kish, G. B. 1973. Stimulus-Seeking and Conservatism. In *The Psychology of Conservatism*, ed. G.D. Wilson. London: Academic Press.

Knight, Kathleen. 2006. "Transformations of the Concept of Ideology in the Twentieth Century." *American Political Science Review* 100(4):619–626.

Mannheim, Karl. 1955. *Ideology and Utopia.* New York: Brace and Company.

Marx, Karl and Friedrich Engels. 1845. *The German Ideology.* Buffalo: Prometheus Books.

McCloskey, Herbert. 1958. "Conservatism and Personality." *American Political Science Review* 52:27–45.

Mendes, W.B., J. Blascovich, S.B. Hunter, B. Lickel and J.T. Jost. 2007. "Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners." *Journal of Personality and Social Psychology* 92:698–716.

Oxley, Douglas R., Kevin B. Smith, John R. Alford, Matthew V. Hibbing, Jennifer L. Miller, Mario Scalora, Peter K. Hatemi and John R. Hibbing. 2008. "Political Attitudes Vary with Physiological Traits." *Science* 321:1667.

Poole, Keith. 2000. "Non-Parametric Unfolding of Binary Choice Data." *Political Analysis* 8:211–237.

Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* New York: Oxford University Press.

Rawls, John. 1971. *A Theory of Justice.* Cambridge: Belknap.

Rawls, John. 1995. *Political Liberalism.* New York: Columbia University Press.

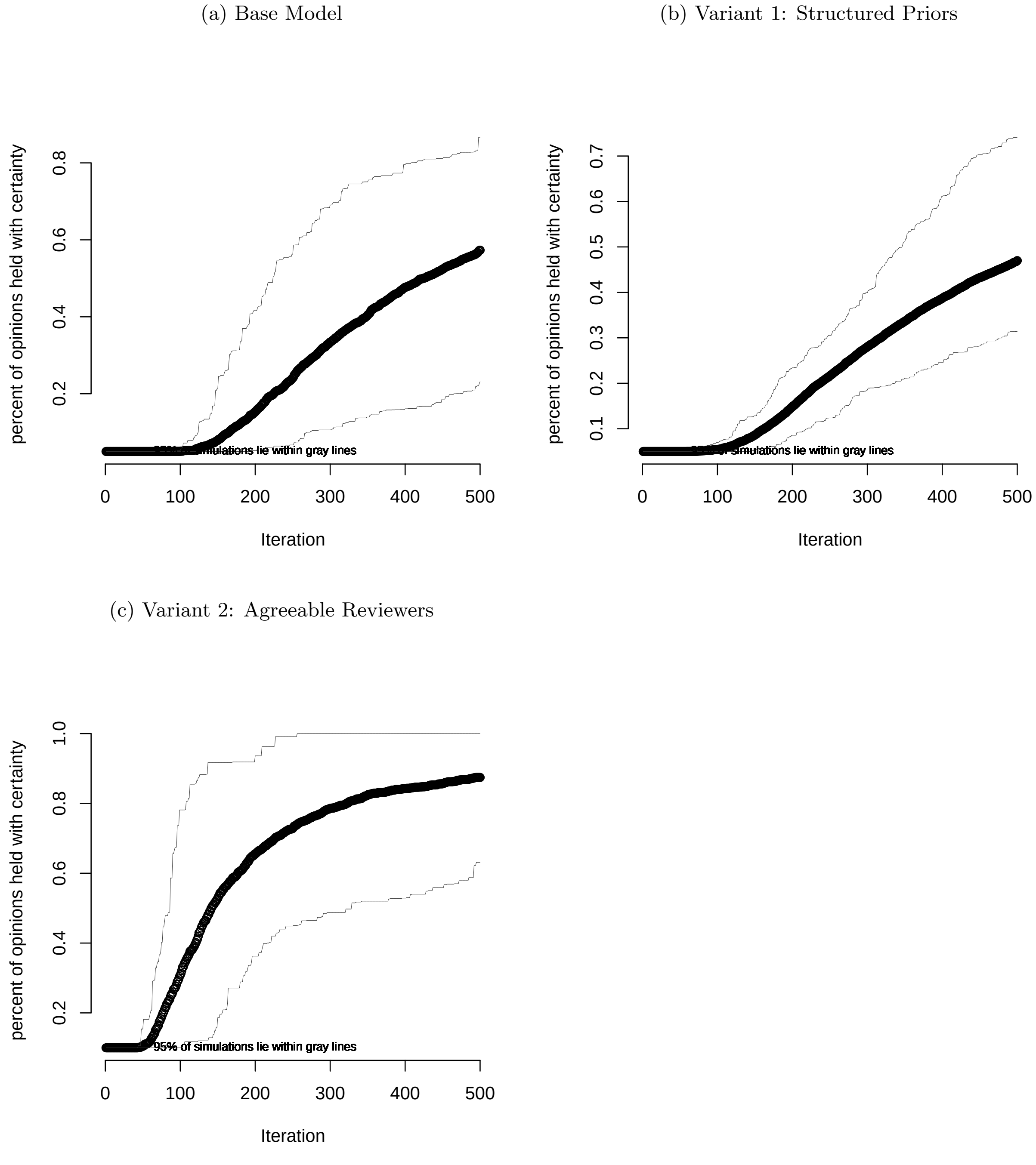
Rawls, John. 2001. *Justice as Fairness: A Restatement.* Cambridge: Belknap Press.

Schwartz, Thomas. 1989. Why Parties? Research memorandum.

Sears, David O., Richard R. Lau, Tom R. Tyler and Harris M. Jr. Allen. 1980. "Self-Interest vs. Symbolic Politics in Policy Attitudes and Presidential Voting." *American Political Science Review* 74(3):670–684.

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

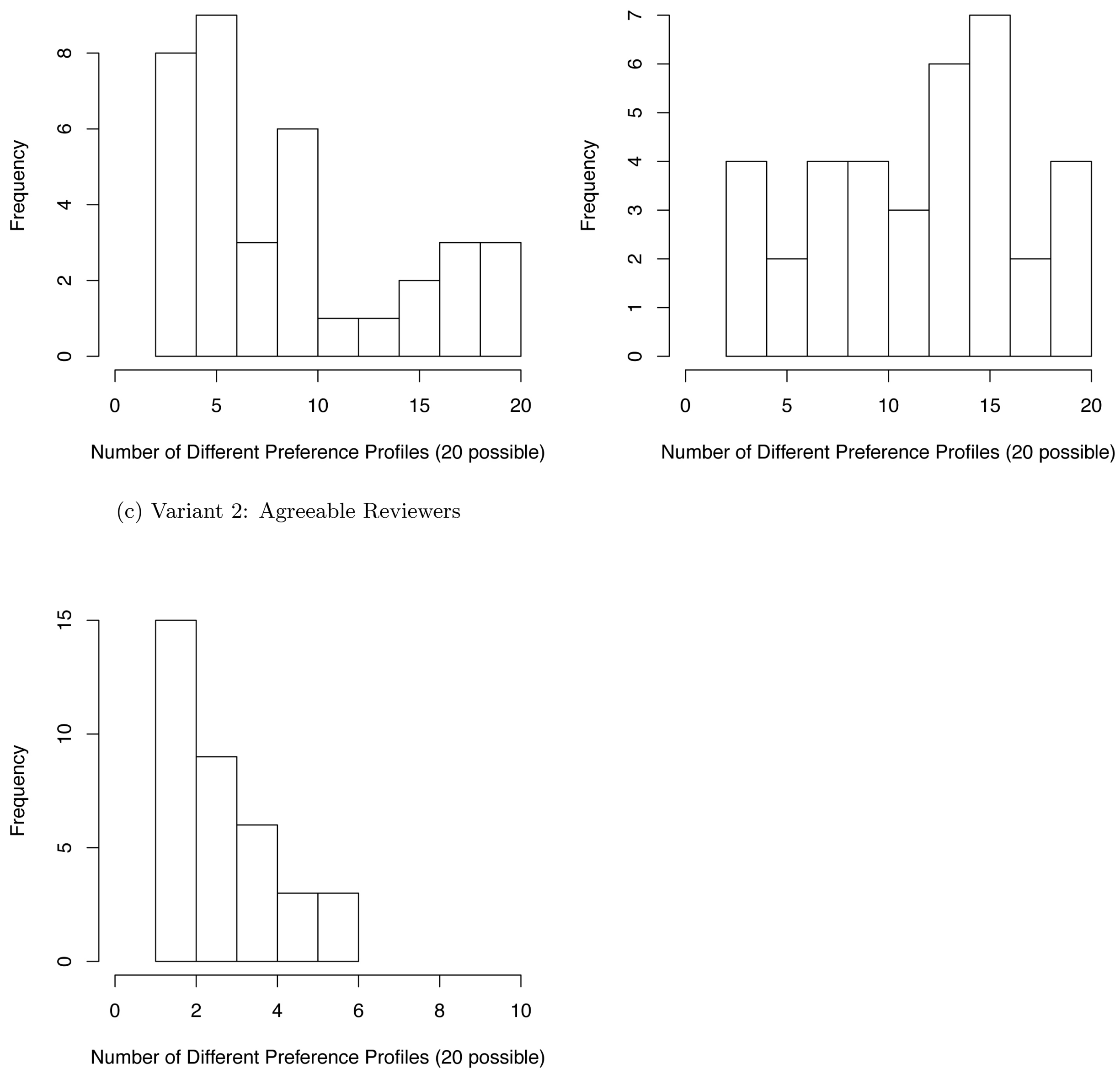
Figure 1: Average Number of Positions Held With Certainty Over Time


(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers

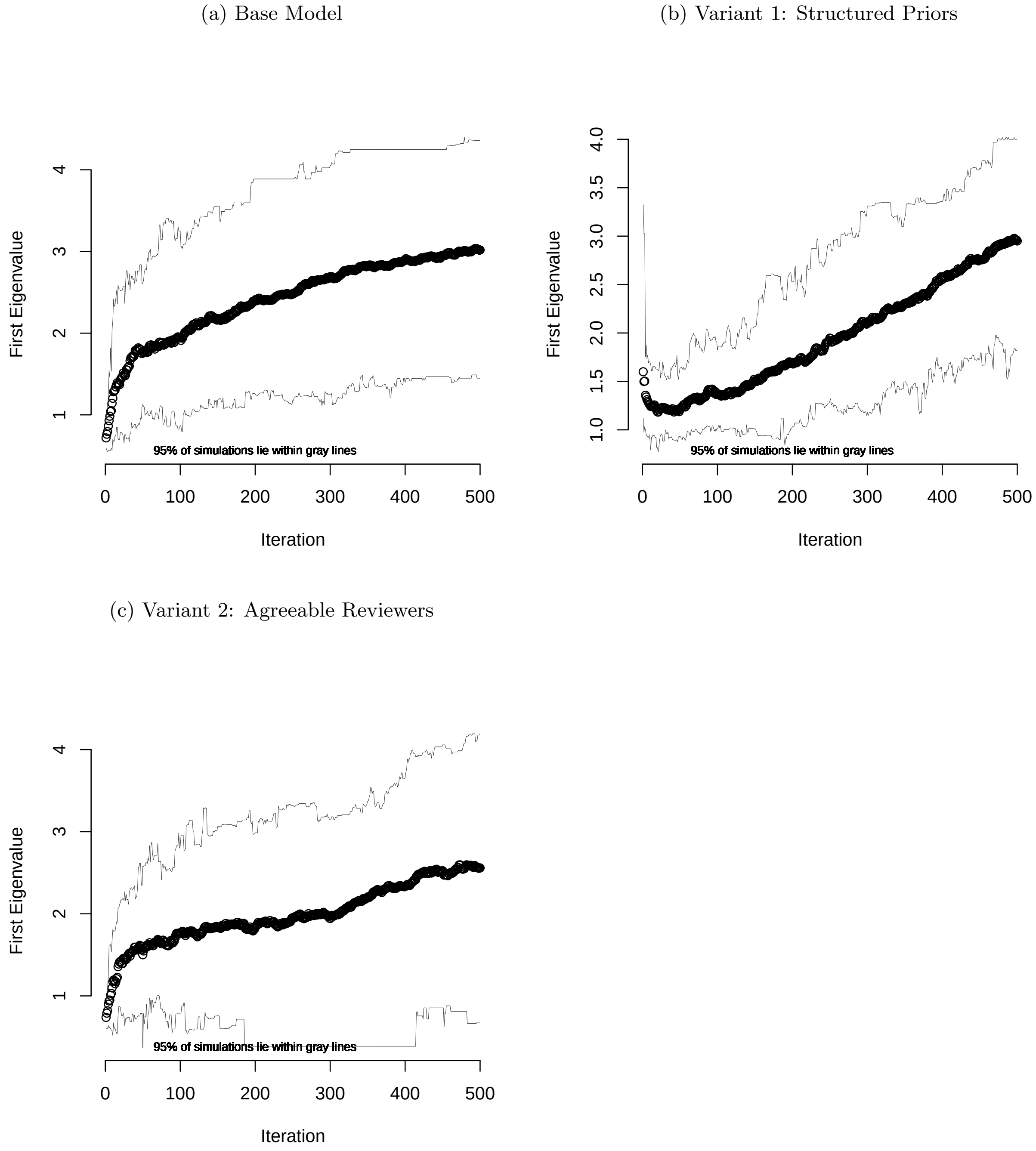
Figure 2: Distribution of Unique Preference Profile at End of 500 Iterations


(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers

Figure 3: Average First Eigenvalue Over Time

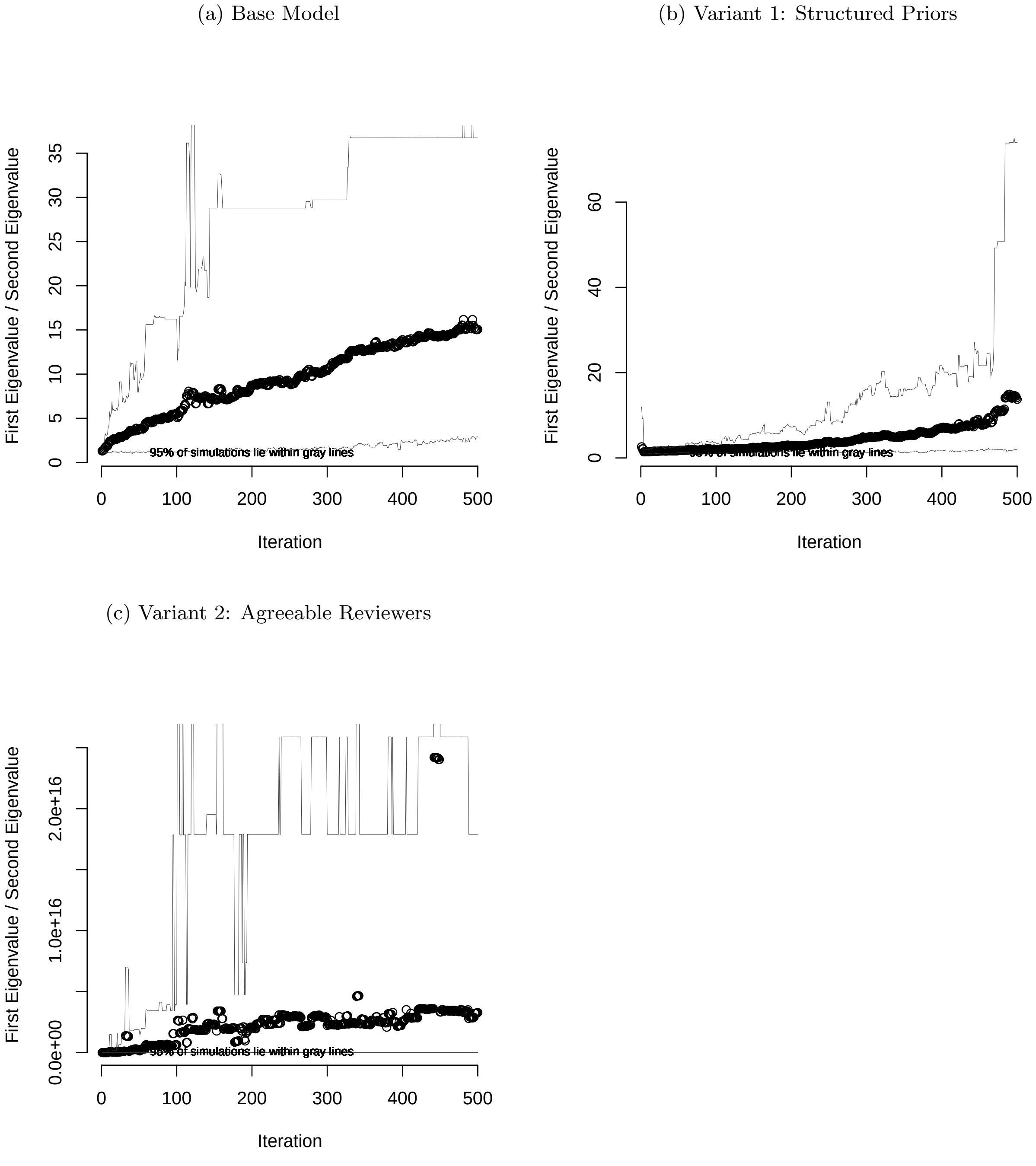


(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers

Figure 4: Average Ratio of First Two Eigenvalues Over Time

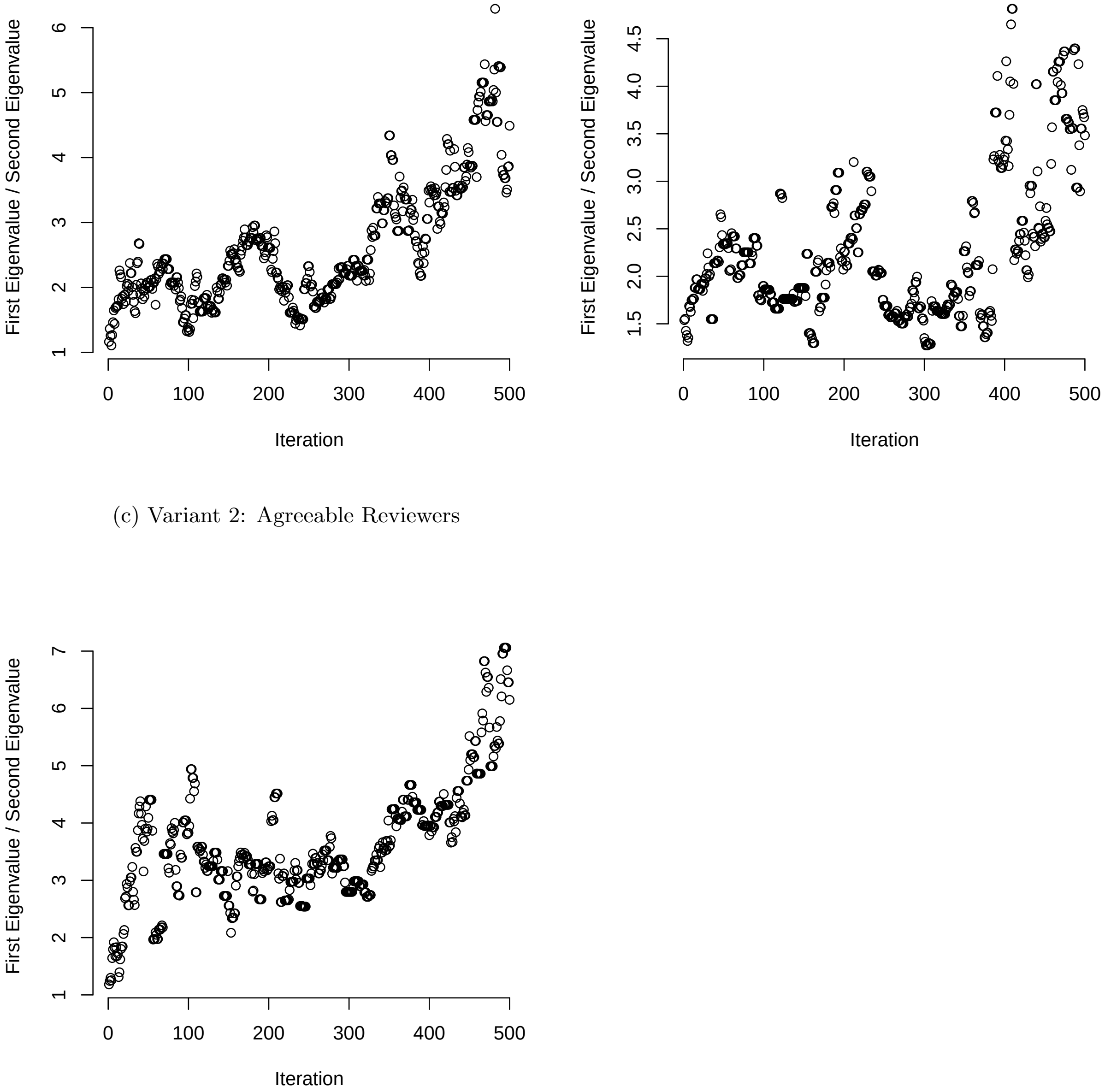


(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers

Figure 5: Ratio of First Two Eigenvalues for One Simulation

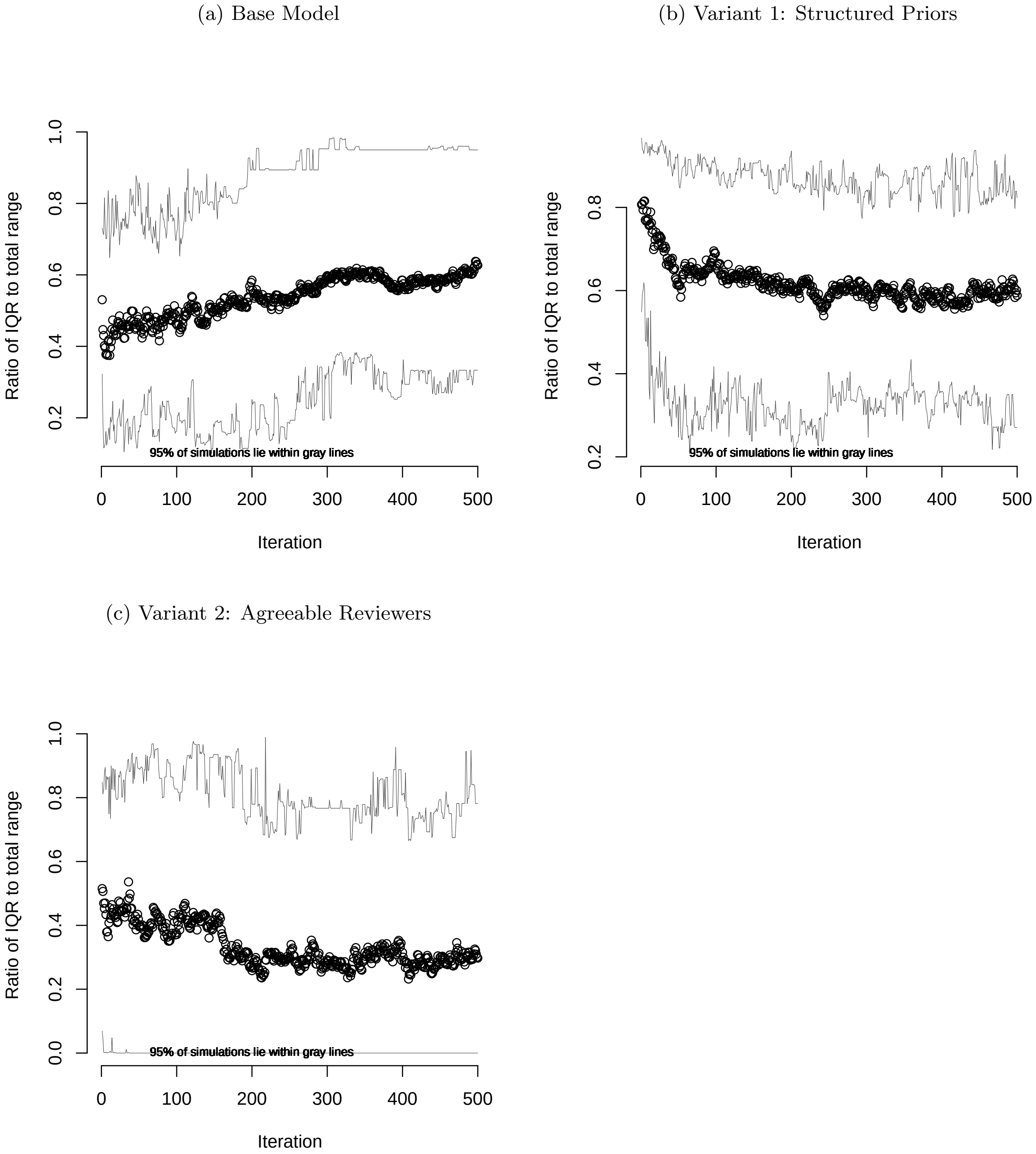


(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers

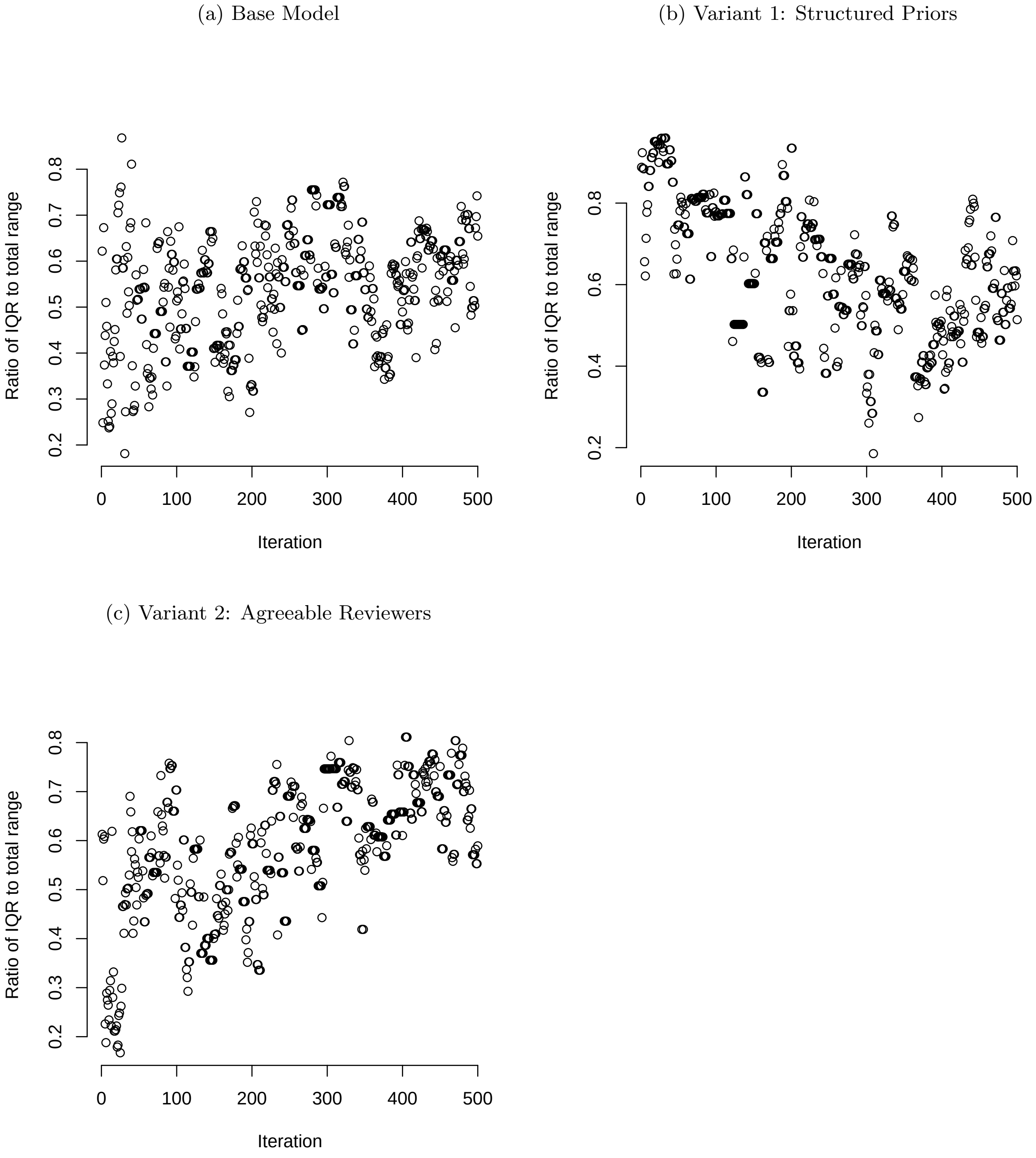
Figure 6: Average Ratio of IQR to Range of Ideal Points Over Time


(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers

Figure 7: Ratio of of IQR to Range of Ideal Points for One Simulation



(a) Base Model

(b) Variant 1: Structured Priors

(c) Variant 2: Agreeable Reviewers