Leadership, Followership, and Beliefs About the World: Theory and Experiment*

Eric S. Dickson
Assistant Professor
Department of Politics and
Center for Experimental Social Science
New York University

January 13, 2008

Behavioral literature in political science and psychology suggests that factual beliefs about the world often vary across different social groups, including by partisan affiliation. This paper explores potential microfoundations for this regularity by analyzing interactions between a group leader and her "followers." A game-theoretic framework is developed, in which a leader with private information about the state of the world sends a message about it to her followers. In the model, the leader is best off when followers successfully coordinate their actions. Because followers have preferences over coordination outcomes that are aligned in some states of the world, but not in others, leaders sometimes have incentives to misrepresent the state of the world in order to make coordination more likely. Two novel refinements, "Leadership-Correlated Equilibrium" and "Belief-Based Followership Equilibrium," explicate distinct mechanisms through which a leader's messages about the world could potentially coordinate the actions of fully-rational followers. The intuitions behind these refinements are then tested in a laboratory experiment. Results from the experimental games suggest that leaders do frequently misrepresent the state of the world to followers when it is in their interests to do so; that leaders' strategically-chosen messages about the state of the world are highly effective in coordinating group members' actions; and that leaders' messages strongly influence followers' beliefs about the state of the world even when Bayesian followers would not find the messages to be credible as statements of fact. Followers appear not fully to account for leaders' strategic incentives to misrepresent the state of the world in forming their posterior beliefs, even though the experimental elicitation mechanism offers substantial monetary rewards for accuracy. This finding suggests behavioral foundations for the empirical regularity that members of different social groups may have different factual beliefs about the world.

*Very preliminary draft; apologies for typos or infelicitous passages. Comments and suggestions very welcome: eric.dickson@nyu.edu

1 Introduction

> "Leadership: the art of getting someone else to do something you want done because he wants to do it."
>
> -Dwight Eisenhower

"Leadership" is among the most prominent themes in contemporary political discourse – yet it is badly undertheorized within contemporary political science. Dwight Eisenhower's implicit claim that good leaders are effective at least in part because they are able to shape others' preferences is both intuitively compelling and widespread in the public sphere. Yet, from a theoretical perspective, it is far from clear through what causal mechanism political or other group leaders might be expected to exert such influence.

It is both natural and useful to understand leadership as one, albeit key, aspect of group dynamics more generally. Of course, decades-old intuitions within both behavioral political science and social psychology suggest that group membership causally shapes not only the way in which individuals see the world generally, but also the way in which they come to understand their own interests specifically in the course of complicated social and political interactions. Within political science, partisan affiliation is arguably the most heavily-studied kind of group membership. The classic formulation of Campbell *et al* (1960) that party identification raises a "perceptual screen through which the individual tends to see what is favorable to his partisan orientation" is suggestive that group membership has a causal effect because it influences cognition, and ultimately individual beliefs, about the political world. This line of argument has continued to the present day; contemporary research, employing more rigorous methodological techniques, suggests that individual beliefs concerning even basic factual information about politics and the social world can be strongly polarized by partisanship (e.g. Bartels 2002, Achen and Bartels 2006).

The causal pathways that may give rise to such relationships are poorly understood; however, it is natural to suppose that an important role may be played by communications between group elites and group members. Such communications, of course, may take on a variety of different forms and

fulfill a number of different roles. Certain aspects of political leadership have been well-captured by game-theoretic models in which "expert" leaders have private information about the state of the world, and can sometimes strategically reveal or dissemble about this information in the context of a signalling game with fully-rational voters (e.g., Canes-Wrone, Herron, and Shotts 2001). At the same time, classical signalling models seem less than ideally equipped to confront the kinds of striking polarization in factual beliefs by partisanship detailed in the literature. While deception can (within limits) be sustainable in signalling-game equilibria, and while correlations between partisan affiliation and exposure to specific media outlets also doubtless play a role, the apparently robust nature of such behavioral phenomena cry out for further theoretical explication.

This paper takes as its starting point the centrality of coordination problems to group life (Schelling 1960). Within the political arena, elections, wars, and other kinds of conflicts are often contests between competing partisan, national, ethnic, or other groups which must be able to engage in meaningfully coordinated action in order to meet their objectives. Previous research within a number of disciplines has argued that leadership can be one instrumentally useful means by which collectives can achieve coordinated outcomes more efficiently, from game-theoretic (Calvert 1992), lab-experimental (Wilson and Rhodes 1997), and evolutionary (Van Vugt 2006) viewpoints.

This paper presents an analytic framework explicitly modelling one process, novel in the literature, by which group leaders might be able to induce coordinated action among their followers. A game-theoretic framework is developed, in which a leader with private information about the state of the world sends a message about it to her followers. In the model, the leader is best off when followers successfully coordinate their actions. Because followers have preferences over coordination outcomes that are aligned in some states of the world, but not in others, leaders sometimes have incentives to misrepresent the state of the world in order to make coordination more likely.

Two novel refinements, "Leadership-Correlated Equilibrium" and "Belief-Based Followership Equilibrium," explicate distinct mechanisms through which a leader's messages about the world could potentially coordinate the actions of fully-rational followers. In the first approach, Leadership-Correlated

Equilibrium, Followers use messages from a Leader as simple correlation devices guiding their actions. Under this approach, the way in which messages affect Followers' beliefs, or the "content" of messages as factual statements about the world more generally, are effectively decoupled from Followers' choices of actions. In the second approach, Belief-Based Followership Equilibrium, Followers use messages from a Leader as potentially informative signals to be used in updating their beliefs about the state of the world. Under this approach, Leaders' messages can affect Followers' prospects for coordination by shaping Followers' perceptions of the underlying nature of their interaction with one another, which depend on their beliefs about the state of the world.

The intuitions behind these refinements are then tested in a laboratory experiment. Results from the experimental games suggest that leaders do frequently misrepresent the state of the world to followers when it is in their interests to do so; that leaders' strategically-chosen messages about the state of the world are highly effective in coordinating group members' actions; and that leaders' messages strongly influence followers' beliefs about the state of the world even when Bayesian followers would not find the messages to be credible as statements of fact. Followers appear not fully to account for leaders' strategic incentives to misrepresent the state of the world in forming their posterior beliefs, even though the experimental elicitation mechanism offers substantial monetary rewards for accuracy. This finding suggests behavioral foundations for the empirical regularity that members of different social groups may have different factual beliefs about the world.

The paper contains five further sections. Section 2 presents a formal framework for analyzing leadership and coordination. Section 3 motivates and offers definitions for Leadership-Correlated Equilibrium and Belief-Based Followership Equilibrium, as well as presenting some benchmark theoretical results. Section 4 describes the experimental design as well as the conditions under which the experiments were carried out. Section 5 presents and discusses results from the experiment. Finally, Section 6 offers further discussion, sets the paper in a broader context, and concludes.

2 A Framework for Analyzing Leadership and Coordination

This section describes a "Leadership and Coordination game," in which a Leader sends a message about the state of the world to two Followers, who subsequently choose alternatives in a coordination game.[1] The ultimate payoffs of all actors – Leader and Followers alike – are determined by Followers' choices.[2]

In the game, the state of the world $\omega \in \Omega = \{1, 2, 3\}$ is determined by a random draw from a probability distribution for which $\omega = n$ with probability $\rho_n \forall n \in \{1, 2, 3\}$, where $\sum_n \rho_n = 1$. The prior probabilities ρ_n are common knowledge. In addition, the Leader knows the state of the world ω; Followers only know the prior probabilities.

First, the Leader sends a message $m \in \{1, 2, 3\}$ about the state of the world to both group members. This message is costless, and its contents may or may not correspond to the true state of the world.

Then, each of two Followers, F_A and F_B, must choose either of two options, A or B. Because the interaction between these Followers takes on the form of a coordination game, each Follower prefers to choose A (B) given that her opponent will choose A (B). However, the precise nature of the payoffs depends on the state of the world, which from the perspective of Followers is *ex ante* uncertain.

Followers F_A and F_B have preferences over coordination outcomes that are aligned in some states of the world, but in conflict in other states of the world. Specifically, as summarized in the payoff matrices below, for F_A a coordination outcome (A, A) yields a payoff $\pi = 1$ when $\omega \in \{1, 2\}$ and a payoff $\pi = \mu \in (0, 1)$ when $\omega = 3$; an outcome (B, B) instead yields a payoff $\pi = \mu$ when $\omega \in \{1, 2\}$ and a payoff $\pi = 1$ when $\omega = 3$. In contrast, for F_B, a coordination outcome (A, A) yields a payoff

[1] For clarity of prose, throughout the Leader is referred to using female pronouns, while the Followers are referred to using male pronouns.

[2] This analytic framework, involving a leader's attempts to influence the dynamics of a two-person group, is of course highly stylized; however, the underlying intuitions about the role of leadership in settings that place a premium on group coordination are meant to extend to interactions within larger groups as well. However, the literal-minded reader can also interpret the model more strictly, in terms of a group leader's attempts to affect social interactions within pairs of group members who may come to interact during the course of group life. In either case, it is easy to imagine that both leaders as well as group members may, in many situations, have an interest in successfully coordinated action – even in the presence of within-group, across-individual differences as to *which* coordination outcome ought to serve as a locus of group action.

$\pi = 1$ when $\omega = 1$ and a payoff $\pi = \mu$ when $\omega \in \{2, 3\}$; an outcome (B, B) instead yields a payoff $\pi = \mu$ when $\omega = 1$ and a payoff $\pi = 1$ when $\omega =\in \{2, 3\}$. An outcome involving a coordination failure, i.e. (A, B) or (B, A), yields payoff $\pi = 0$ for both Followers in every state of the world.

Payoff Matrix for State of the World $\omega = 1$.

.	F_B: A	F_B: B
F_A: A	$1, 1$	$0, 0$
F_A: B	$0, 0$	μ, μ

Payoff Matrix for State of the World $\omega = 2$.

.	F_B: A	F_B: B
F_A: A	$1, \mu$	$0, 0$
F_A: B	$0, 0$	$\mu, 1$

Payoff Matrix for State of the World $\omega = 3$.

.	F_B: A	F_B: B
F_A: A	μ, μ	$0, 0$
F_A: B	$0, 0$	$1, 1$

The Leader also has preferences over outcomes that depend on the Leader's type, $\theta \in \{L_A, L_B\}$. A leader of type L_A receives the same payoffs for every outcome, in every state of the world, as Follower F_A; in the same way, a leader of type L_B receives the same payoffs from outcomes as Follower F_B. The Leader knows her own type, while Followers share a common prior belief that $\theta = L_A$ with probability $q \in [0, 1]$ but that $\theta = L_B$ with complementary probability $1 - q$.

This preference structure reflects an intuitive account of group life. Successful coordination is essential for achieving many group ends; yet, members of a group may sometimes agree, and sometimes disagree, about the coordinated outcome they would most like to achieve. In the framework here, the two Followers agree about *which* coordination outcome they individually prefer in two of the three states of the world ($\omega = \{1, 3\}$); these will be referred to as *Agreement States.* In the third state of the world ($\omega = 2$), however, there is disagreement; $\omega = 2$ will be referred to as the *Disagreement State.* The potential for disagreement alongside the possibility of agreement, all within the context of an intertwined destiny as modeled here by a coordination game, is typical of the interactions of political or other social groups. As an example, members of a given political party may share an interest in,

and need for, coordination for the purposes of winning power. However, different group members may exhibit differing tendencies; for example, some members of a given political party may be more inclined to be foreign policy doves, while others possess more hawkish inclinations. Such distinctions indicate the potential for disagreement under certain circumstances – while the distinct factions may potentially find themselves in complete agreement under others. Hawks may favor bold intervention in some states of the world for which doves would find such action imprudent; however, in a state of the world following a direct attack against the country by a menacing threat, or in a state of the world where any present threat is considered by all sides to be minor and more amenable to negotiations, group members may be united in their preferences over courses of action.

3 Two Fully-Rational Conceptions of Leadership in Equilibrium

This section describes two distinct, and complementary, approaches to modeling behavior using Perfect Bayesian Equilibria of the Leadership and Coordination Game. In the first approach, Leadership-Correlated Equilibrium, Followers use messages from a Leader as simple correlation devices guiding their actions. Under this approach, the way in which messages affect Followers' beliefs, or the "content" of messages as factual statements about the world more generally, are effectively decoupled from Followers' choices of actions. In the second approach, Belief-Based Followership Equilibrium, Followers use messages from a Leader as potentially informative signals to be used in updating their beliefs about the state of the world. Under this approach, Leaders' messages can affect Followers' prospects for coordination by shaping Followers' perceptions of the underlying nature of their interaction with one another, which depend on their beliefs about the state of the world.

Both Leadership-Correlated Equilibrium and Belief-Based Followership Equilibrium can be thought of as particular refinements of Perfect Bayesian Equilibrium in the context of the Leadership and Coordination game. In any state of the world, there are three equilibria in the coordination subgame following receipt of the Leader's message: coordination on (A, A), coordination on (B, B), and a Mixed-Strategy Nash Equilibrium. Because of this, there are a very large number of Perfect Bayesian

Equilibria in the Leadership and Coordination game as a whole. As such, the Leadership-Correlated and Belief-Based Followership Equilibrium concepts are far from exhaustive in characterizing potential equilibria in the game as a whole. However, these equilibrium concepts are useful in explicating some of the mechanisms through which leadership can potentially help group members overcome coordination problems of different kinds. In addition, both equilibrium concepts are supported by intuitive, behaviorally plausible refinement assumptions, which is not the case for some of the other Perfect Bayesian equilibria of the Leadership and Coordination game.

Before proceeding to these equilibrium concepts, it is useful to define several terms that will enhance clarity of discussion.

Definition 1. Self-Interested Agreement State. *Suppose that in state of the world ω, a Leader's most-preferred outcome is X. Then the Leader's* self-interested Agreement State ω^* *is the state in which X is both Followers' most-preferred outcome.*

According to this definition, a Leader of type L_A has self-interested Agreement State $\omega^* = 1$ when the true state is $\omega = \{1, 2\}$, but $\omega^* = 3$ when $\omega = 3$; a Leader of type L_B has self-interested Agreement State $\omega^* = 1$ when $\omega = 1$, but $\omega^* = 3$ when $\omega = \{2, 3\}$. Intuitively, Leaders can potentially benefit when Followers believe the state of the world to be the Leader's self-interested Agreement State, ω^*, because both Followers will perceive a preference for coordination on the Leader's most-preferred outcome.

Definition 2. Self-Interested Message. A Leader's message m is a *self-interested message* if $m = \omega^*$.

Thus, a Leader's message will be referred to as "self-interested" if the message communicates to Followers that the Leader's self-interested Agreement State is in fact the state of the world. Note that, according to this definition, it is possible for a self-interested message either to be truthful (when $\omega = \omega^*$) or untruthful (when $\omega \neq \omega^*$).

"Leadership-Correlated Equilibrium"

One pathway through which Leaders' communications may coordinate Followers is simply through providing focal points around which Followers may rally. In certain settings, in the absence of a message from a group Leader, neither A nor B may constitute a focal alternative for Followers. When this is the case, it seems natural to suppose that the mixed-strategy equilibrium of the coordination subgame will be selected. A Leader's message to her Followers may allow for the selection of a more efficient equilibrium, in which coordination failures never occur, by giving a focal property either to a choice of A or to a choice of B.

In the framework presented above, Leaders send messages m whose manifest content is about the state of the world, ω. The preceding paragraph suggests however that such messages may influence Followers' prospects for coordination in a way that is effectively decoupled from actors' actual beliefs about the state of the world. A message $m = 1$, for example, may bestow a focal property on alternative A, because the coordination outcome (A, A) is Pareto-optimal in state of the world $\omega = 1$. Potentially, such a message could serve as a purely rhetorical device that aids coordination independent of actors' beliefs about the state of the world.

Of course, in many circumstances, fully-rational Followers can nonetheless learn something about the state of the world from Leaders' messages, even when such messages merely serve as coordination devices. Within the context of any separating equilibrium, different Leader types can be expected to send different messages in at least some states of the world. However, although Followers in such a context could use Bayes' Rule to learn more about the state of the world, such updating of beliefs will not be directly relevant to Followers' behavior in the Leadership-Correlated Equilibrium model; Followers will simply choose whatever alternative is made focal by Leaders' messages.

In the abstract, it is naturally the case that any number of assumptions could be made about *which* alternatives are made focal by *which* messages about the world. However, it seems behaviorally reasonable to make a refining assumption that focality is aligned with the natural-language meaning of messages received, in the following sense. In a context in which Followers' actions cue off of Leaders'

speech, a message $m = 1$, evoking $\omega = 1$, would naturally call to mind the action A, because outcome (A, A) is Pareto-optimal under $\omega = 1$. The definition of Leadership-Correlated Equilibrium includes this refining assumption:

Definition 3. A *Leadership-Correlated Equilibrium* is a Perfect Bayesian Equilibrium of the Leadership and Coordination game in which:

(i) Followers who receive an Agreement State message choose the alternative on which coordination would be Pareto-optimal in that Agreement State, that is, A for $m = 1$ and B for $m = 3$;

(ii) Followers who receive a Disagreement State message play the Mixed Strategy Nash Equilibrium of the coordination subgame because neither A nor B is focal; and

(iii) Leaders choose their optimal message, given this profile of Follower responses.

Given this definition, it is straightforward to establish the following Proposition:

Proposition 1. The Leadership and Coordination game has a unique Leadership-Correlated Equilibrium, in which (1) the Leader sends her self-interested message and (2) Followers coordinate on the leader's preferred outcome.

Sketch of Proof. By the definition of a Leadership-Correlated Equilibrium, both Followers choose A upon receiving $m = 1$; B upon receiving $m = 3$; and both play the Mixed Strategy Nash Equilibrium upon receiving $m = 2$. Therefore, any Leader can obtain her most-preferred outcome by choosing her self-interested message. Leaders have no incentive to deviate from this strategy profile because they always receive their most-preferred outcome; an individual Follower has no incentive to deviate from this strategy profile because by construction the strategy profile is Nash in every coordination subgame. ■

Thus, Leaders' messages about the state of the world have the potential to "organize" Followers' behavior even when Followers' beliefs about the world are decoupled from their actions. From a technological standpoint, this coordination mechanism requires only that Agreement State messages can bestow a focal property on one particular alternative, an assumption that seems reasonable in many behavioral contexts. Nonetheless, the Leadership-Correlated Equilibrium model is quite thin as

a model of political communication. In many contexts, it seems natural to suppose that Followers' beliefs about the world are not irrelevant to their behavior in equilibrium. To develop this intuition, the next section describes a second equilibrium concept for the Leadership and Coordination game.

"Belief-Based Followership Equilibrium"

In the "Belief-Based Followership" model, Followers' beliefs about the state of the world do have relevance to Followers' choices, and the ultimate prospects for successful coordination. Under this equilibrium concept, Leaders' messages influence Follower behavior through a more subtle mechanism; rather than directly cueing Followers to choose one particular alternative over another, Leaders' messages affect the beliefs of fully-rational Followers, and through this the Followers' perceptions of their own interests, thereby affecting the shape of Followers' coordination game interactions with one another.

Consider the strategy profile in which, as in the previous section, a Leader always sends her self-interested message. That is, a Leader of type L_A sends $m = 1$ when $\omega = \{1, 2\}$, but $m = 3$ when $\omega = 3$, while a Leader of type L_B sends $m = 1$ when $\omega = 1$, but $m = 3$ when $\omega = \{2, 3\}$. Given this strategy profile for the different Leader types, and prior beliefs about the state of the world defined by the ρ_i, a Bayesian Follower will form posterior beliefs $\overline{\rho_i}(m = j)$ about the state of the world given a message $m = j \in \{1, 3\}$:

$\overline{\rho_1}(m = 1) = prob(\omega = 1|m = 1) = \frac{\rho_1}{\rho_1 + \rho_2 q}$

$\overline{\rho_2}(m = 1) = prob(\omega = 2|m = 1) = \frac{\rho_2 q}{\rho_1 + \rho_2 q}$

$\overline{\rho_3}(m = 1) = prob(\omega = 3|m = 1) = 0$

$\overline{\rho_1}(m = 3) = prob(\omega = 1|m = 3) = 0$

$\overline{\rho_2}(m = 3) = prob(\omega = 2|m = 3) = \frac{\rho_2(1-q)}{\rho_3 + \rho_2(1-q)}$

$\overline{\rho_3}(m = 3) = prob(\omega = 3|m = 3) = \frac{\rho_3}{\rho_3 + \rho_2(1-q)}$

These posterior beliefs will be common knowledge for both Followers, given the structure of the game, since it is commonly known that both have access to the same information.

Followers' perceived preferences over the set of coordination outcomes, $\{(A, A), (B, B)\}$, will of

course be a function of their beliefs about the state of the world, since Follower payoffs are state-dependent. For concreteness, suppose that in a given equilibrium context, Leaders always send their self-interested messages, as above. If Follower F_B receives message $m = 1$, and has posterior belief $\overline{\rho_1} > \frac{1}{2}$, then in expectation he believes that (A, A) would be his most-preferred outcome, rather than (B, B).

Importantly, Followers' posterior beliefs about the state of the world affect not only their individual beliefs about their privately most-preferred outcome, but also the *kind* of coordination game that both Followers commonly perceive one another to be playing. For instance, if Followers receive message $m = 1$, and share posterior belief $\overline{\rho_1} > \frac{1}{2}$, then in expectation both Followers believe themselves to be best off when the outcome is (A, A). That is, the coordination game that the Followers perceive themselves to be playing is one with a Pareto-dominant Nash equilibrium. On the other hand, if Followers receive message $m = 1$, but share posterior belief $\overline{\rho_2} > \frac{1}{2}$, then in expectation Follower F_A believes himself to be best off when the outcome is (A, A), while Follower F_B believes himself to be best off when the outcome is (B, B), because as described above, Followers' preferences differ when $\omega = 2$. That is, the coordination game that the Followers perceive themselves to be playing is a Battle-of-the-Sexes game, without a Pareto-dominant Nash equilibrium.

While Followers' payoffs are of course determined by the payoff matrix corresponding to the true state of the world ω, Followers' behavior in the coordination game can naturally be expected to depend on the game they *perceive* themselves to be playing, given their posterior beliefs about the state of the world. Because some coordination "problems" can reasonably be thought of as more-easily solvable than others, and because the nature of the coordination game that is perceived depends on Followers' posterior beliefs about the world, it is easy to see in this framework that Leaders with an interest in coordinating Followers' actions will sometimes have an incentive to try to influence Followers' beliefs in one or another direction, to the extent this can be done in equilibrium, consistent with Bayes' Rule.

That some coordination problems are more easily solved than others is reflected in the following

assumptions about the outcomes of play in the different coordination games that players perceive.[3] If, given their shared posterior belief about the state of the world, both Followers believe in expectation that the same outcome is best for each of them, then it will be assumed that both Followers coordinate on that outcome with probability 1. In other words, a common belief that (A, A) is best for both Followers leads to the selection of the (A, A) equilibrium in the coordination subgame; similarly, a common belief that (B, B) is best for both Followers leads to the selection of the (B, B) equilibrium in the coordination subgame. This assumption is meant to reflect that this perceived coordination game is relatively easy to solve because the players' perfectly aligned incentives produce a strongly focal desirable outcome.

Now suppose that the two Followers perceive their preferences over the two coordination outcomes to differ. This will be the case if Follower F_A perceives (A, A) to be the best outcome for himself at the same time that Follower F_B perceives (B, B) to be the best outcome for himself. In such a case it is assumed that the group members will fail to coordinate with probability $\mathcal{F} \in (0, 1)$ – that is, the outcome of play will be one of the un-coordinated outcomes, (A, B) or (B, A), with strictly positive probability. Alternatively, they may in fact coordinate successfully on (A, A) or (B, B), each with probability $\frac{1-\mathcal{F}}{2}$ (this equiprobability reflecting the symmetry of the perceived game form). The introduction of a nonzero probability of coordination failure when group members' perceived interests differ is meant to reflect that coordination is harder to achieve when incentives are imperfectly aligned than when players are in complete agreement. For the present purposes of defining the equilibrium concept, the value $\mathcal{F}$ will be taken to be rate of coordination failure when Followers play the Mixed-Strategy Nash equilibrium. Of course, in specific real-world settings, group members may be able to coordinate more often than the rate implied by the MSNE, though still not all the time, if they have access to some other partially-effective correlating device. In such instances $\mathcal{F}$ can be modelled differently, as appropriate; as a marker for such potential generalizations, $\mathcal{F}$ is left general in the notation.

[3] These assumptions will be further justified momentarily.

As the results rely on the assumption that some coordination problems are more easily solved than others, it is worthwhile to reflect a moment on why this assumption might be a reasonable one. The perceived payoff matrices when group members believe themselves to be in agreement are coordination games with a focal, Pareto-dominant equilibrium, whereas the perceived payoff matrices when group members believe themselves not to be in agreement are battle of the sexes games. The relative ease with which coordination can be achieved in any concrete instance will of course vary depending upon a variety of factors, including the degree to which players are free to communicate; however, the intuition that coordination will on average be more easily achieved in coordination games with a Pareto-dominant equilibrium than in battle of the sexes games seems persuasive over a wide variety of conditions. If players must make choices without communication, a coordination game with a Pareto-dominant equilibrium will have a clear focal point that battle of the sexes games will lack, and as such, all other things equal, coordination can be expected to be easier in the former case than in the latter. This might be stated in another way. Both games have two pure strategy Nash equilibria and one mixed-strategy Nash equilibrium. Because there is a good reason for players to focus on one particular pure strategy equilibrium in the coordination game with a Pareto-dominant equilibrium, but there is no good reason for players to focus on one particular pure strategy equilibrium in the battle of the sexes game, they are more likely to play the mixed-strategy equilibrium in the battle of the sexes game - and this equilibrium involves a positive probability of coordination failure. Under conditions of free communication, the basic intuition remains intact. When group members believe their interests to be aligned, it is natural to expect that actors will quickly agree on the Pareto-optimal equilibrium virtually all of the time. Meanwhile, the battle-of-the-sexes game would pose a potentially non-trivial bargaining problem even under conditions of full communication. The experimental literature on bargaining provides some insight here: "One of the clearest experimental results, which also accords well with field data, is that a nonnegligible frequency of disagreements [bargaining failures] is a characteristic of bargaining in virtually all kinds of environments." (Roth 1995) This is the case even when it is clear that it is in all actors' interests to forge a deal: "While this would be unsurprising if it occurred only in situations

that presented the bargainers with no mutually profitable agreements, most of the evidence suggests that disagreements and costly delays are pervasive even when it is evident that there are gains to be had from agreement." (Roth 1995)

Because players may or may not perceive the state of the world correctly, the coordination game that players understand themselves to be engaged in may or may not correspond to the true payoff matrix reflecting the actual state of the world. While the probabilities of group members' possible coordination outcomes are determined based on the payoff matrix that they perceive to be relevant, their ultimate payoffs are derived from the true payoff matrix corresponding to the actual state of the world. As a clarifying example, suppose that the state of the world is $\omega = 2$, but imagine that both Followers instead have posterior belief $\overline{\rho_1} = 1$. In this case, both Followers perceive a preference for (A, A) as the best-possible outcome. According to the coordination-game assumptions, this implies that (A, A) will be the outcome. Group members' payoffs are then determined according to the payoff matrix corresponding to $\omega = 2$ – that is, F_A receives 1 while F_B receives mu.

This discussion motivates the following definition:[4]

Definition 4. A *Belief-Based Followership Equilibrium* is a Perfect Bayesian Equilibrium of the Leadership and Coordination game in which:

(i) Followers with posterior beliefs $\overline{\rho_1} > \frac{1}{2}$ and $\overline{\rho_2} = 1 - \overline{\rho_1}$ play A, resulting in coordination on (A,A);

(ii) Followers with posterior beliefs $\overline{\rho_3} > \frac{1}{2}$ and $\overline{\rho_2} = 1 - \overline{\rho_3}$ play B, resulting in coordination on (B,B);

(iii) Followers with posterior beliefs $\overline{\rho_2} > \frac{1}{2}$ and either $\overline{\rho_1} = 1 - \overline{\rho_2}$ or $\overline{\rho_3} = 1 - \overline{\rho_2}$ play the Mixed Strategy Nash Equilibrium of the coordination subgame because neither A nor B is focal;

(iv) Upon receiving any off-the-equilibrium path message m_k, Followers have degenerate posterior

[4]The discussion of Belief-Based Followership Equilibrium is focused around the strategy profile in which Leaders of either type always send self-interested messages. As will be seen in a moment, this strategy profile does not always support an equilibrium of the game. Additionally, when the strategy profile does support an equilibrium of the game, this equilibrium is not necessarily unique. The next draft of the paper will tie up these loose ends, either by further refining the definition of BBFE to ensure uniqueness when there is existence, or by characterizing the set of equilibria under the current definition of BBFE.

belief $\overline{\rho_k} = 1$; and

(v) Leaders choose their optimal message, given this profile of Follower responses.

It will prove useful later to define one additional term inspired by Definition 4:

Definition 5. A message $m = 1$ is *credible as a statement of fact* if $\overline{\rho_1} > \frac{1}{2}$ and $\overline{\rho_2} = 1 - \overline{\rho_1}$. Similarly, a message $m = 3$ is *credible as a statement of fact* if $\overline{\rho_3} > \frac{1}{2}$ and $\overline{\rho_2} = 1 - \overline{\rho_3}$.

Given these definitions, it is straightforward to establish the following Proposition:

Proposition 2. *Consider a strategy profile in which a Leader of either type always sends her self-interested message in every state of the world. Then:*

Case I. *When* $\frac{\rho_1}{\rho_1+\rho_2 q} > \frac{1}{2}$ *and* $\frac{\rho_3}{\rho_3+\rho_2(1-q)} > \frac{1}{2}$*, this strategy profile is supported in a Belief-Based Followership Equilibrium, all Leaders' messages are credible as statements of fact, and Followers coordinate on the Leader's preferred alternative;*

Case II. *When* $\frac{\rho_1}{\rho_1+\rho_2 q} < \frac{1}{2}$ *and* $\frac{\rho_3}{\rho_3+\rho_2(1-q)} < \frac{1}{2}$*, this strategy profile is supported in a Belief-Based Followership Equilibrium, any Leaders' messages* $m = 1$ *or* $m = 3$ *are not credible as statements of fact, and Followers play the Mixed Strategy Nash Equilibrium;*

Case III. *Suppose* $\frac{\rho_1}{\rho_1+\rho_2 q} < \frac{1}{2}$ *and* $\frac{\rho_3}{\rho_3+\rho_2(1-q)} > \frac{1}{2}$*. When* $\mu > \frac{(1-\mathcal{F})(1+\mu)}{2}$*, this strategy profile is not supported in a Belief-Based Followership Equilibrium, because a Leader of either type would have an incentive to deviate from* $m = 1$ *to* $m = 3$ *when* $\omega = 1$*. However, when* $\mu < \frac{(1-\mathcal{F})(1+\mu)}{2}$*, this strategy profile is supported in a Belief-Based Followership Equilibrium;*

Case IV. *Suppose* $\frac{\rho_3}{\rho_3+\rho_2(1-q)} < \frac{1}{2}$ *and* $\frac{\rho_1}{\rho_1+\rho_2 q} > \frac{1}{2}$*. When* $\mu > \frac{(1-\mathcal{F})(1+\mu)}{2}$*, this strategy profile is not supported in a Belief-Based Followership Equilibrium, because a Leader of either type would have an incentive to deviate from* $m = 3$ *to* $m = 1$ *when* $\omega = 3$*. However, when* $\mu < \frac{(1-\mathcal{F})(1+\mu)}{2}$*, this strategy profile is supported in a Belief-Based Followership Equilibrium.*

Proof Sketch. Given the strategy profile in which a Leader of either type always sends her self-interested message in every state of the world, consider the incentives of Leader types L_A and L_B. $\underline{\omega = 1}$. The strategy profile assigns both Leader types to send $m = 1$. When $\overline{\rho_1} = \frac{\rho_1}{\rho_1+\rho_2 q} > \frac{1}{2}$, the outcome is (A, A) by the definition of Belief-Based Followership Equilibrium. This yields the maximum

payoff 1 to both L_A and L_B so neither has an incentive to deviate. When instead $\overline{\rho_1} = \frac{\rho_1}{\rho_1+\rho_2 q} < \frac{1}{2}$, then $\overline{\rho_2} > \frac{1}{2}$; the outcome is (A, A) with probability $\frac{1-\mathcal{F}}{2}$, (B, B) with probability $\frac{1-\mathcal{F}}{2}$, and a coordination failure with probability $\mathcal{F}$. Thus, the expected payoff to either Leader type is $\frac{(1-\mathcal{F})(1+\mu)}{2}$. Deviation to $m = 2$ yields posterior belief $\overline{\rho_2} = 1$, by the off-the-equilibrium-path assumption in the definition of BBFE, which yields the same distribution of outcomes, so there is no incentive to deviate. Finally, deviation to $m = 3$ yields posterior belief $\overline{\rho_3} = \frac{\rho_3}{\rho_3+\rho_2(1-q)}$. If $\overline{\rho_3} = \frac{\rho_3}{\rho_3+\rho_2(1-q)} < \frac{1}{2}$, deviating to $m = 3$ again leads to an unchanged expected payoff $\frac{(1-\mathcal{F})(1+\mu)}{2}$, while if $\overline{\rho_3} = \frac{\rho_3}{\rho_3+\rho_2(1-q)} > \frac{1}{2}$, the outcome is (B, B) and the payoff to either Leader type is μ. Thus, each of the Leader types has an incentive to deviate if and only if both $\frac{\rho_1}{\rho_1+\rho_2 q} < \frac{1}{2}$ and $\frac{\rho_3}{\rho_3+\rho_2(1-q)} > \frac{1}{2}$ while $\mu > \frac{(1-\mathcal{F})(1+\mu)}{2}$. $\underline{\omega = 3}$. By a symmetric argument to that for $\omega = 1$, each of the Leader types has an incentive to deviate if and only if both $\frac{\rho_3}{\rho_3+\rho_2(1-q)} < \frac{1}{2}$ and $\frac{\rho_1}{\rho_1+\rho_2 q} > \frac{1}{2}$ while $\mu > \frac{(1-\mathcal{F})(1+\mu)}{2}$. $\underline{\omega = 2}$. Leader type L_A (L_B) has the same preferences over outcomes, and according to the strategy profile sends the same message, as when $\omega = 1$ ($\omega = 3$). Hence, the incentives for deviation by L_A (L_B) from $m = 1$ ($m = 3$) when $\omega = 2$ are the same as the incentives for deviation by both Leader types when $\omega = 1$ ($\omega = 3$).

Taken together, these results, along with the coordination-subgame equilibrium selection assumptions in the definition of BBFE, prove the Proposition. ■

* FIGURE 1 ABOUT HERE *

Figure 1 offers a graphical depiction of the conditions on the ρ's for the individual Cases in the Proposition. The axes depict the values of ρ_1 and ρ_2; the value of ρ_3 is implicit because $\rho_3 = 1-\rho_1-\rho_2$. With the help of the Figure, the intuition behind Proposition 2 is quite straightforward. In Case I, the prior probability ρ_2 of the Disagreement State $\omega = 2$ is quite low relative to the prior probabilities of the individual Agreement States. Under these conditions, a Leader's self-interested message, indicating one or the other of the Agreement States, will seem credible as a statement of fact to Followers. As a result, Leaders are able to obtain their first-best outcome, even when the true state of the world is a Disagreement State.

In contrast, the prior probability of the Disagreement State is quite high relative to the prior

probabilities of the individual Agreement States in Case II. When this is true, a Leader's self-interested message will *not* seem credible as a statement of fact to Followers, who in the context of the Leader's strategy profile will believe that the Disagreement State is most likely regardless of the message received. Because a Leader cannot improve her position by deviating to another message, it is an equilibrium for Leaders always to send their self-interested message; however, these messages are not persuasive as they are in Case I, and Followers play the Mixed Strategy Nash Equilibrium of the coordination subgame.

Finally, it is not an equilibrium for a Leader always to send his or her self-interested message in Cases III or IV when μ is sufficiently high. Consider Case III, in which the prior probability ρ_1 of $\omega = 1$ is low relative to the prior probabilities of both other states. In these circumstances, a Leader's self-interested message $m = 1$ would not be credible as a statement of fact, given the strategy profile; Followers would believe the Disagreement State to be more likely than $\omega = 1$. However, the prior probability of the other Agreement State, $\omega = 3$, is sufficiently higher that a Leader's self-interested message $m = 3$ *would* be credible as a statement of fact. Because of this, Leaders have an incentive to deviate from $m = 1$ to $m = 3$, in order to ensure coordination (even if it is not on their most-preferred outcome), so long as μ is sufficiently large.

4 Experimental Instantiation

The above theoretical exposition has developed several key intuitions. Central among these are that communications from leaders can facilitate coordination among group members, and that leaders may sometimes have incentives to misrepresent the state of the world in these communications. If group members are fully rational, certain messages from leaders will be understood as credible, while others will not be. Whether a message's credibility affects its usefulness as a coordinating device is a point of distinction between the Leadership-Correlated and Belief-Based Followership equilibrium concepts.

A laboratory experiment was conducted as a means of exploring the intuitions underlying these theoretical approaches. Five experimental sessions were conducted in a social science lab at a large

American university. The 96 subjects, each of whom took part in one session only, interacted anonymously via networked computers; the experiments were programmed and conducted with the software z-Tree (Fischbacher 1999). Participants, almost all of whom were undergraduates from around the university, signed up via a web-based recruitment system that draws on a large, pre-existing pool of potential subjects. Subjects were not recruited from the author's courses. After giving informed consent according to standard human subjects protocols, subjects received written instructions that were subsequently read aloud in order to promote understanding and induce common knowledge of the experimental scenario.

At the beginning of each session, subjects were randomly assigned to a group of three people, consisting of a randomly-assigned "Group Speaker" and two "Group Members"; these role labels were thought to be more neutral than Leader and Follower, their analogues in the theoretical exposition.[5] Group and role assignments remained fixed over 15 periods of interaction.

Each period consisted of one play of the Leadership and Coordination stage game, followed by a "Bonus Question" for Followers that served as an incentive-compatible mechanism for eliciting their beliefs about the state of the world. The game-theoretic structure of the Leadership and Coordination stage game used in the lab was identical to that of the framework in Section 2. Subjects earned "tokens," convertible into dollars at the end of the experiment (at a rate of 15 tokens = US$1), according to the outcomes of play (and, for Followers, their private "Bonus Question" responses); overall payoffs were equal to the sum of payoffs from each of the 15 periods, plus a US$5 show-up fee.

The stage games in each of the 15 periods were formally independent – that is, all random-variable quantities such as the state of the world and Leaders' preferences were re-drawn independently in every period, in a way to be described in more detail. In addition, a variety of design features were intended to minimize the extent to which subjects could condition behavior on the previous period's outcome of

[5]The Appendix contains a sample set of instructions to subjects, including an extensive series of screenshots showing the computer interface, that offers a complete depiction of the way the experiment was framed for participants. The description in this section places terminology from the experimental scenario in quotation marks where this differs from the theoretical exposition; for continuity, however, the analysis is presented in the same terms as the earlier discussion in the paper.

play. The coordination game alternatives (A and B in the theoretical exposition) were referred to using labels that varied from one period to the next; the elements in a given pair of alternatives were places, people, or things belonging to the same category (e.g., Cleveland/Cincinnati, Dandelion/Daffodil, Schooner/Sloop, etc.). It was intended that neither label in any pair seem particularly salient relative to the other, and each subject saw any given pair of labels only once. To further inhibit conditioning on previous periods, the states of the world ($\omega = \{1, 2, 3\}$ in the theoretical exposition) were referred to using the names of colors as labels; for every group, in every period, three color names were drawn at random for this purpose from a set of ten.[6] Various features of the on-screen interface were also randomized; for example, it was known to subjects that alternatives and states of the world were both presented in an order that had been randomized for every subject. Finally, to ensure that neither Follower's preferences became focal relative to the other's, each Follower was referred to on his own screen as "you," while his counterpart Follower was referred to as "your counterpart." On Leaders' screens, the two Followers were referred to as "Group Member 1" and "Group Member 2" for clarity, but Followers never observed these labels, and it was known to all subjects that these labels were randomly reassigned from one period to the next.

Subjects were aware that each group's state of the world was in fact randomly drawn based on the commonly-known "likelihoods" (prior probabilities) shown to the group. These likelihoods (ρ_1, ρ_2, and ρ_3, with $\rho_1 + \rho_2 + \rho_3 = 1$) took on positive integer percentage point values (e.g., 9%, 53%, etc.) that were themselves randomly and independently drawn for every group, in every period, according to a procedure that was not described to subjects. This procedure modestly oversampled the region of the probability simplex corresponding to Case II (using one round of rejection sampling; 27.3% of the likelihood triples drawn came from this region, which covers 16.7% of the probability simplex) in order to increase Followers' exposure to circumstances in which Leaders had clear incentives to misrepresent the state of the world. Aside from this oversampling (and rounding considerations), the likelihood triples were uniform draws from the probability simplex.

[6]The colors were White, Gray, Black, Yellow, Orange, Red, Brown, Green, Blue, and Purple.

All five sessions were run with $q = 0.5$; each Leader's type was independently drawn in every period, so the Leader was equally likely to share each of her Followers' preferences in every period. Two sessions (with 21 subjects each) were run with payoff parameter $\mu = 0.8$, while two sessions (with 21, 21, and 12 subjects respectively) were run with $\mu = 0.2$; because patterns of Leader and Follower behavior were highly similar across these two conditions, all analyses below pool data from all five sessions. For both Leaders and Followers, the payoff for coordination on one's most-preferred alternative was 10 "tokens" (= US$0.667), while a failure to coordinate yielded no payoff.

In each period, Followers answered a "Bonus Question" after choosing a coordination game alternative, but before receiving any feedback about the outcome of play. This "Bonus Question" simply asked Followers to guess what the state of the world for the period actually had been; a correct answer earned 10 additional "tokens," while an incorrect answer earned none. Thus, the "Bonus Question" served as an incentive-compatible mechanism for eliciting Followers' posterior beliefs about the most likely state of the world. Because the 10 "token" (= US$0.667) reward for a correct answer was equal in magnitude to the maximum possible payoff from the play of the game, subjects were well-motivated by lab-experimental standards to learn the state of the world as well as possible.[7] It was common knowledge that Leaders' payoffs were not affected by Followers' Bonus Question answers; in place of Bonus Question payoffs, Leaders simply received a flat payment of 5 extra tokens per period.

As feedback at the end of each period, Leaders and Followers were informed which alternatives had been chosen by the two Followers. Leaders, who of course knew the state of the world, therefore learned their precise payoffs for the period. Followers, who were at no point told the state of the world, did not learn their precise payoffs, either for game play or for their "Bonus Question" response. They

[7]Typical psychology studies exhibiting "biases" in decision making do not offer subjects concrete incentives based on their responses; as a result, such studies are treated skeptically by many political scientists and economists. Any demonstration that some kind of "biased" behavior remains intact even when subjects' responses are rewarded with a financial incentive makes a stronger case for that bias as a robust phenomenon. Of course, it is reasonable to imagine that the presence of such motivation leads subjects to attend to the experimental task differently than they would in the absence of the motivation, which may pose a problem for inference or for an experiment's external validity in some contexts. However, foreshadowing the results, this feature of the design actually strengthens inference in the present experiment; subjects exhibit a particular kind of bias in the lab even though they have additional incentives to be unbiased, relative to the incentives likely faced in real-world contexts.

simply observed whether or not coordination had been achieved during that period.

This limited feedback to Followers constituted an important part of the experimental design. Obviously, it was desirable to have Followers participate in a number of periods, over which they might be exposed to instances of different theoretical cases, gain familiarity with the experimental interface, and develop deeper insights into the dynamics of the experimental game. A demonstration of a "bias" in belief formation based on one period only would clearly be less convincing than a demonstration of a persistence of some "bias" over repeated interactions. Further, by interacting with a given Leader over a substantial number of experimental periods, Followers potentially could learn from patterns of Leader behavior over time; for example, from a Bayesian perspective it could easily be inferred that a Leader who never sent a Disagreement State message over many periods was in fact lying about the state of the world at least some of the time. Allowing Followers to learn about Leader behavior in the context of such an ongoing relationship arguably parallels important dynamics in the real world, where leaders speak to group members about many issues and many facets of the world over time. More substantial feedback to Followers – for example, revealing the true state of the world at the end of every period – would short-circuit such dynamics by making Leaders' honesty, or lack of it, unrealistically transparent; although there are exceptions, in real-world politics there is seldom an external arbiter of truth about the state of the world that a group member would have access to or find definitively credible.

5 Experimental Results

Leaders' Messages

A key insight behind the theoretical model is that leaders may sometimes have incentives strategically to misrepresent the world in order to facilitate coordination among group members. The data indicates that subjects in the role of Leader overwhelmingly perceive, and respond to, these incentives in the context of the experiment.

Experimental Result 1. *In an Agreement State, Leaders nearly always accurately report the state of the world. But in a Disagreement State, Leaders overwhelmingly misrepresent the state of the world*

by claiming it to be an Agreement State.

Consistent with theoretical expectations from both models, Leaders were highly averse to sending Disagreement State messages, regardless of the true state of the world. During those periods conducted in an Agreement State, Leaders' messages accurately reported the state of the world an overwhelming 93.5% of the time (272/291); only 1.7% of the time (5/291) did they send a message corresponding to the Disagreement State. In sharp contrast, during those periods conducted in a Disagreement State, Leaders' messages accurately reported the state of the world only 22.8% of the time (43/189). Instead, in a substantial majority of cases (61.4%, 116/189), Leaders in a Disagreement State sent their "self-interested" message, indicating that "Agreement State" in which the Leader's most-preferred option would be Pareto-optimal; leaders sent a message corresponding to the other "Agreement State" in 15.9% of cases (30/189). The tendency of Leaders to misrepresent Disagreement States as Agreement States increased somewhat over time as experimental sessions progressed.[8]

The 48 instances in which Leaders' messages announced the Disagreement State were distributed very unequally across subjects in the role of Leader. 15 of the 32 experimental Leaders *never* sent such a message, while 5 Leaders sent such a message only once. Of the remaining 43 instances, fully 20 were due to only three Leaders, who each sent such a message at least six times; the other 23 were distributed among 9 different Leaders.

* TABLE 1 about here *

Of course, the Belief-Based Followership equilibrium concept makes different predictions across different theoretical Cases, depending on the *ex ante* likelihoods of different states of the world. Table 1 disaggregates Leaders' messages by Case.[9]

Under Cases I and II, both the Leadership-Correlated and Belief-Based Followership concepts pre-

[8]Leaders honestly reported a Disagreement State 29.9% (29/97) of the time in the first 7 periods but 15.2% (14/92) of the time in the last 8 periods; this difference in proportions is significant ($p = 0.016$, two-tailed test). They instead indicated that Agreement State in which the Leader's most-preferred option would be Pareto-optimal 53.6% (52/97) of the time in the first 7 periods but 69.6% (64/92) of the time in the last 8 periods ($p = 0.024$, two-tailed test).

[9]Aggregate totals slightly exceed the sum of Case-by-Case totals because a small number of draws from the probability simplex fell directly onto a boundary between Cases. Data from periods with these draws is included in the aggregate totals, but not tallied as belonging to any of the four Cases.

dict that Leaders will accurately report the state of the world in Agreement States[10], but will misrepresent it in Disagreement States by sending their self-interested message instead. The bulk of Leaders' experimental messages are in fact broadly consistent with this pattern, although Leaders do send honest messages reporting Disagreement States between a fifth and a quarter of the time. In a small number of instances, Leaders instead report the Agreement State in which their less-preferred outcome is likely to result; in two-thirds of these instances (20/30), this "other" Agreement State is more *ex ante* likely than the Leader's "self-interested" Agreement State, raising the possibility that some Leaders may at the margins take into account the plausibility of their messages to Followers.

Under Cases III and IV, Leaders continue to send their self-interested message in Leadership-Correlated equilibria, but have incentives instead to indicate the "other" Agreement State according to the Belief-Based Followership concept. In fact, Leaders in Agreement States almost always honestly report the state of the world under Cases III and IV, consistent with the incentives perceived by Leaders in Leadership-Correlated equilibria. As in Cases I and II, Leaders honestly report Disagreement States between a fifth and a quarter of the time, and sometimes report the "other" Agreement State, most prominently in Disagreement State periods when their "preferred" Agreement State is unlikely.

Followers' Actions

A further insight of the model is that Followers will be able to use Leaders' messages as coordinating devices, thereby achieving much higher rates of successful coordination than would be possible in the absence of such messages. The data displayed in Table 2 indicates that coordination rates are indeed extremely high, contingent on receipt of an Agreement State message from the Leader.

* Table 2 About Here *

Experimental Result 2. *When Leaders send an Agreement State message, Followers successfully coordinate at very high rates.*

Specifically, when a Leader's message indicated an Agreement State, individual Followers chose the

[10] As elsewhere, in the current draft, for BBFE this refers to the specific strategy profile explicated in the theoretical section.

alternative upon which coordination would be Pareto optimal in that state an overwhelming 96.9% (837/864) of the time. This pattern of choices led to an overall coordination rate of 93.8% (405/432) conditional on an Agreement State message. 22 of the 32 groups *never* failed to coordinate upon receipt of such a message; 6 groups accounted for fully 22 of the 27 failures (which occurred either three or four times each).

The practical utility of Agreement State messages for coordination is clear when comparing the 93.8% figure with the overall coordination rate of 52.1% (25/48) achieved following Disagreement State messages, which of course do not make one coordination alternative focal in the way that Agreement State messages do. A two-sample test of proportions unsurprisingly indicates that the difference between these coordination rates is highly significant ($p << 0.0001$, two-tailed).

Experimental Result 3. *Coordination rates remain high when Leaders' messages are not credible as statements of fact, but they are statistically significantly lower than coordination rates when Leaders' messages are credible.*

At first glance, a coordination rate of 93.8% would seem to leave relatively little variation to analyze. However, a closer look reveals that rates of coordination do appear to vary as leaders' messages appear more or less plausible as statements of fact, given the prior probabilities of different states of the world. First, given an Agreement State message, coordination failures occur in Case II (10.0%, 11/110) at more than twice the rate they occur in Case I (4.5%, 9/199); the difference in proportions is significant ($p = 0.060$, two-tailed). Of course, Leaders' self-interested messages are credible in Case I, but not in Case II, suggesting that messages may lose some of their ability to coordinate Followers when their content is less believable as fact. A second notable finding is that messages indicating the *ex ante more* likely Agreement State induced coordination fully 96.9% (281/290) of the time, while messages indicating the *ex ante less* likely Agreement State did so only 86.6% (116/134) of the time. The fourfold difference in rates of coordination failure implicit in these two proportions is highly significant ($p = 0.0001$, two-tailed test). Again, the evidence suggests that Leaders' messages lose some of their coordinating power when they appear less likely to be truthful.

These results offer something of a mixed message for the two theoretical accounts in the previous section. On one hand, Leaders' Agreement State messages are highly effective in inducing coordination even when the messages should not be credible as actual statements about the world, more in accordance with Leadership-Correlated Equilibrium than with the Belief-Based Followership approach. On the other hand, Leaders' messages are indeed less effective in coordinating Followers when these messages appear to be less credible as statements of fact, a result anticipated by Belief-Based Followership, but not by Leadership-Correlated Equilibrium.

Followers' Beliefs about the State of the World

The above evidence suggests that Leaders act on incentives strategically to misreport the state of the world, and that Leaders' messages can be highly effective in coordinating Followers' actions. One central question remains: do Followers form posterior beliefs about the state of the world in the same way that Bayesians would, factoring in Leaders' strategic incentives to dissemble? Or are Followers' beliefs influenced by Leaders' messages in a different way, unpredicted by the Bayesian theory? Table 3 addresses this question by summarizing Followers' guesses about the state of the world, disaggregated by the type of message received and by Case.

* Table 3 about here *

Experimental Result 4. *In Case I, Leaders' Agreement State messages are credible, and Followers almost always believe that these messages accurately portray the state of the world.*

In Case I, given Leaders' strategies (either in the Belief-Based approach or in Leadership-Correlated Equilibrium), Agreement State messages are credible. As such, Followers who act as Bayesian agents would rationally guess that the message is an accurate depiction of the state of the world. In fact, Followers do this an overwhelming 90.2% (359/398) of the time.

Experimental Result 5. *In Case II, Leaders' Agreement State messages are not credible, but Followers nonetheless believe about 40% of the time that these messages accurately portray the state of the world.*

Given Leaders' strategies in Case II, whether in the equilibrium discussed in the Belief-Based Followership exposition or in the unique Leadership-Correlated Equilibrium, Bayesian Followers should place posterior probability of at least $\frac{1}{2}$ on the Disagreement State regardless of Leaders' messages, because the Disagreement State is so likely *ex ante*. Yet, Followers who receive an Agreement State message in Case II guess that the Disagreement State is the state of the world only 56.8% (125/220) of the time. Fully 42.7% (94/220) of the time, Followers instead guess that the Leader's message accurately communicated the state of the world, even though the message was not credible as a statement of fact from a Bayesian perspective.

This result is striking. In the experimental scenario, Leaders' incentives to misrepresent a Disagreement State as an Agreement State are quite transparent – certainly more so than in corresponding real-world settings. Leaders only receive positive payoffs when Followers coordinate successfully, and coordination occurs at vastly higher rates after Agreement State messages have been sent. Yet, it appears that Followers often fail to account for Leaders' strategic incentives when interpreting the meaning of these messages. That is, Followers often appear to take Leaders' messages as true statements of fact even when they are highly likely to be false. As a result, in the aggregate, Followers' beliefs about the state of the world are much more strongly influenced by Leaders' messages than would be the case if Followers behaved as Bayesian agents.

The remainder of this section considers the robustness of, and the interpretation of, this key finding. A first piece of evidence for Followers' mistaken beliefs as a robust behavioral phenomenon is that Followers exhibit little aggregate improvement as they gain experience and listen to successive messages from Leaders. Indeed, in Case II, Followers who receive a non-credible Agreement State message guess that the state of the world is the Disagreement State 54.5% (48/88) of the time in the first 7 periods, and 58.8% (77/131) of the time in the last 8 periods; the difference between these proportions is statistically insignificant ($p = 0.53$, two-tailed).

At the individual level, there was considerable variation in the extent to which different Followers accounted for Leaders' strategic incentives in formulating beliefs about the state of the world. Upon

receiving an Agreement State message in Case II, 23 Followers always guessed the Disagreement State to be the true state of the world, consistent with Bayesian inference, while 24 Followers never guessed the Disagreement State to be the true state of the world. The remaining 13 Followers who were exposed to an Agreement State message in Case II split their guesses between Agreement and Disagreement States in different periods.

Of course, the Bayesian inferences about the state of the world that are described above for Cases I and II assume that Leaders truthfully reveal the state of the world in Agreement States, but send Agreement State messages in a Disagreement State. While most Leaders' messages were self-interested, a minority were not. In evaluating the rationality of Followers' beliefs about the world, it is important to account for the potential effects of such deviant messages on the beliefs of even Bayesian-rational Followers, who might reasonably conclude (for example) that a Leader who has previously sent a Disagreement State message may place some value on being honest that is not part of the model, which could lead to different equilibrium strategies for the Leader.

As noted above, messages deviating from the theoretical predictions were distributed very unevenly across Leaders. The simplest way to explore the effects of deviant messages on Followers' beliefs about the world is simply to split the data involving guesses that follow receipt of a Case II Agreement State message into two subsamples – first, all guesses made by Followers who had *never* up to that point seen a Disagreement State message, and second, all other guesses, made after at least one such exposure. Under such circumstances, 60.1% (83/138) of guesses in the first subsample indicated the Disagreement State, compared to 51.2% (42/82) in the second subsample. The difference between these two proportions does approach marginal statistical significance by conventional standards under a one-tailed test ($p = 0.198$, two-tailed test), but perhaps more to the point, the 60.1% figure from the "uncontaminated" (by Disagreement State messages) subsample is not meaningfully different substantively from the overall 56.8% rate reported above. A similar result – 61.9% (78/126) – is obtained when restricting the analysis only to those Followers whose Leader *never at any point* delivered a Disagreement State message.

These results are particularly striking because Followers actually had ample opportunity to learn about the nature of Leader behavior over the course of 15 periods of interaction. Feedback to Followers was limited in the sense that they were not informed of the true state of the world at the end of each period – and could therefore not directly infer that Leaders had misrepresented the world in any *specific* period. However, consider the Followers whose Leader was one of the 15 (out of 32) Leaders who never sent a Disagreement State message. By the middle of period 8, the halfway point of each group's existence, each such Follower would have seen 8 consecutive Agreement State messages from their Leader. This simple fact alone constituted considerable evidence that Leaders in these groups were unlikely to be sending "honest" messages in every period. For the median group in the experiment with such a Leader, there was a mere 0.75% probability that the prior probabilities ρ seen by Followers each period would generate Agreement States in eight successive periods. (The corresponding figure over all 15 periods was 0.0069%.) Yet, Followers in such groups continued to "overbelieve" such messages as statements of fact a considerable fraction of the time.

An alternative approach is to incorporate the empirical frequency with which Leaders send Disagreement State messages into Bayesian inference. In the data, Leaders' messages honestly reveal the Disagreement State 22.8% of the time. Assuming that Leaders do this at random, but in Case II send their self-interested message the other 77.2% of the time, Followers' posterior belief $\overline{\rho_1}(m=1)$ would be:

$$\overline{\rho_1}(m=1) = \frac{prob(m=1|\omega=1)prob(\omega=1)}{prob(m=1|\omega=1)prob(\omega=1)+prob(m=1|\omega=2)prob(\omega=2)} = \frac{\rho_1}{\rho_1+0.772q\rho_2}$$

rather than the equilibrium value $\frac{\rho_1}{\rho_1+0.772q\rho_2}$. Under this new assumption, Bayesian Followers would find a message $m=1$ not to be credible only when the prior probability ρ_2 exceeds a threshold that is larger than the one previously calculated. In terms of Figure 1, this excludes points in the Case II region that are relatively close to the Case boundary lines, leaving only points that lie more deeply within Case II territory. Yet, within this restricted sample, Followers still guess that the Disagreement State is the true state of the world only 62.2% (97/156) of the time, a rate statistically indistinguishable from the full sample ($p = 0.294$, two-tailed).

Taken together, these robustness checks add further strength to the conclusion that Followers fail fully to account for Leaders' strategic incentives in choosing messages a substantial fraction of the time, even after repeated interactions with a given Leader. As a result, Followers' beliefs about the state of the world are biased away from Bayesian judgments, towards Leaders' messages about the world, even when these are not credible as statements of fact.

Finally, it is worthwhile to note that Followers who receive a Disagreement State message almost always guess the state of the world to be the Disagreement State (94.7% or 89/94, aggregated across Cases), quite consistent with the off-the-equilibrium-path assumption made in the theoretical exposition.

6 Discussion and Conclusion

The experimental results presented in the paper offer important insights into the dynamics of leadership and followership, and suggest behavioral microfoundations for across-group differences in factual beliefs about the world. Leaders typically have an interest in enhancing their group's prospects for success, and in many settings, coordinated action by group members is an important factor in achieving good outcomes. This paper offered a theoretical framework in which coordination is easier to achieve in some states of the world than in others, because group members' preferences over outcomes may be aligned in some, but not all, states. This premise suggests that Leaders may sometimes have incentives to misrepresent the state of the world in order to make successful coordination more likely. The paper defined two refinements to Perfect Bayesian Equilibrium as a means of exploring two distinct pathways through which Leaders' messages might induce coordination in Bayesian, fully-rational Followers. However, the experimental results suggest that, in contrast to the Bayesian view, Followers frequently fail to account for Leaders' strategic incentives to misrepresent the state of the world. As a result, Followers' beliefs about the state of the world are biased in the direction of Leaders' strategically-chosen messages. Because different groups have different leaders, similar failures of inference replicated in the real world would lead to polarization in factual beliefs about the world across groups.

Of course, the external validity of laboratory-experimental results to real-world processes is always a matter of interpretation. Yet, in a number of respects, various features of the laboratory environment would seem if anything to make it *less* rather than more likely that Followers would fail to account for Leaders' strategic incentives in formulating beliefs about the world. Subjects in the experiment were directly paid for the accuracy of their guesses about the state of the world, an immediate incentive for clear thinking in forming beliefs that has no ready analogue in mass politics; indeed, Followers could potentially earn as much from these guesses as from the coordination games themselves. Followers were randomly assigned to groups and interacted anonymously with fellow group members, a context presumably largely devoid of the kinds of emotional relationships that individuals may have with groups of which they are members, or with group leaders, in national, partisan, social-identity-group, or other real-world collective contexts. Subjects were given precise, objective information about the probabilities of different states of the world, a feature of the experimental protocol that made these prior probabilities far more accessible and substantially more salient than in nearly any real-world setting of interest. Clear inferences about the actual state of the world, given Leaders' messages, were doubtless easier to make given these features of the experiment than they would have been if such information had been less salient or more vague. And, finally, Leaders' incentives to dissemble by misrepresenting Disagreement States were quite transparent in the experiment, certainly more so than in parallel real-world contexts.

The experimental results offer some support for the intuitions underlying both equilibrium concepts advanced in the paper. Consistent with Leadership-Correlated Equilibrium, leaders usually represent the world as being in an "Agreement State," in which all group members' interests are aligned, and group members achieve high rates of coordination using these messages as correlation devices. But, consistent with Belief-Based Followership Equilibrium, leaders' messages are less effective in coordinating group members' actions when these messages are not credible as statements of fact, although the size of the effect is modest. Of course, the relative salience of different considerations in any simple laboratory environment is unlikely to replicate the relative salience of analogous considerations in

real-world settings, and it would be a mistake to read too much into the finer details of the results. For example, in richer informational environments, leaders' messages may be less directly effective as simple coordination devices than they were in the experiment because some group members may not receive them, or because they may receive other communications that make those messages less focal. In such settings, the mechanism underlying Leadership-Correlated Equilibrium may be less effective, and group members may rely more on their beliefs about the world as guides to action.

Regardless, a larger message from the data is that neither of these fully-rational equilibrium concepts seems likely to offer a complete understanding of leadership and followership, most importantly because neither anticipates group members' failures to account for leaders' strategic incentives when formulating beliefs about the world. For many Followers, Leaders' messages apparently *are* credible as statements of fact about the world, even though by the reckoning of Bayes' Rule they should not be.

While a thorough theoretical understanding remains elusive, this finding is suggestive of new directions for future theoretical development. Imagine a modified version of the Belief-Based Followership model, in which Followers often failed to account for Leaders' strategic incentives in sending messages. Such a model could potentially predict higher coordination rates in Case II than the original Belief-Based Followership model did, because, thanks to Leaders' messages and the way Followers received them, Followers' perceptions of their individual interests would be more highly correlated with one another than would be the case for Bayesian Followers. This thought experiment has the intriguing implication that Bayesian-rational Followers may, under some circumstances, be on average *worse* off than would Followers who more blindly believed what Leaders had to say about the world, because the latter could overcome coordination problems more effectively. Such an implication suggests a tension between individual awareness about the world and group interests; strictly from the perspective of tangible group outcomes, accurate individual perceptions of the state of the world may even be considered a public "bad" under certain circumstances if such perceptions undermine the likelihood of successful coordination. At a more theoretical level, such an implication also suggests a deep, underlying tension between Bayesian rationality and the optimizing project that Bayesian rationality is meant to

serve. The resolution of this tension should be fertile ground for further behavioral and game-theoretic research.

7 References

Achen, Chris and Larry Bartels. 2006. "It Feels Like We're Thinking: The Rationalizing Voter and Electoral Democracy." Princeton University Working Paper.

Bawn, Kathleen. 1999. "Constructing 'Us': Ideology, Coalition Politics, and False Consciousness." *American Journal of Political Science* 43: 303-34.

Calvert, Randall L. 1992. "Leadership and Its Basis in Problems of Social Coordination." *International Political Science Review* 13(1): 7-24.

Campbell, Angus, Philip Converse, Warren Miller, and Donald Stokes. 1960. *The American Voter.* Wiley: New York.

Canes-Wrone, Brandice, Michael Herron, and Ken Shotts. 2001. "Leadership and Pandering: A Theory of Executive Policy-Making." *American Journal of Political Science* 45(3): 532-550.

Green, Donald, Bradley Palmquist, and Eric Schickler. 2002. *Partisan Hearts and Minds.* New Haven: Yale University Press.

Roth, Alvin E. 1995. "Bargaining Experiments," in *The Handbook of Experimental Economics*, John H. Kagel and Alvin E. Roth, eds. Princeton University Press: Princeton.

Schelling, Thomas C. 1960. *The Strategy of Conflict.* Harvard University Press: Cambridge, MA.

Van Vugt, Mark. 2006. "Evolutionary Origins of Leadership and Followership." *Personality and Social Psychology Review* 10(4): 354-371.

Wilson, Rick K. and Carl M. Rhodes. 1997. "Leadership and Credibility in N-Person Coordination Games." *Journal of Conflict Resolution* 41(6): 767-791.

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

Table 1. Leaders' Messages: by Case.

Case I.

In an Agreement State			In a Disagreement State		
preferred (and true) Agreement State	other Agreement State	Disagreement State	preferred Agreement State	other Agreement State	(true) Disagreement State
94.3% (165/175)	4.6% (8/175)	1.1% (2/175)	60.0% (21/35)	14.3% (5/35)	25.7% (9/35)

Case II.

In an Agreement State			In a Disagreement State		
preferred (and true) Agreement State	other Agreement State	Disagreement State	preferred Agreement State	other Agreement State	(true) Disagreement State
88.6% (31/35)	11.4% (4/35)	0% (0/35)	64.6% (62/96)	13.5% (13/96)	21.9% (21/96)

Cases III and IV.

In the Likelier Agreement State			In the Less-Likely Agreement State		
preferred (and true) Agreement State	other Agreement State	Disagreement State	preferred (and true) Agreement State	other Agreement State	Disagreement State
94.4% (67/71)	2.8% (2/71)	2.8% (2/71)	100% (9/9)	0% (0/0)	0% (0/0)

In a Disagreement State when pref'd A.S. is likelier			In a Disagreement State when pref'd A.S. is less-likely		
preferred Agreement State	other Agreement State	(true) Disagreement State	preferred Agreement State	other Agreement State	(true) Disagreement State
71.4% (15/21)	9.5% (2/21)	19.0% (4/21)	44.1% (15/34)	29.4% (10/34)	26.5% (9/34)

Table 2. Coordination Rates: by Case.

	Coordination Rate: Message Indicates Agreement State	Coordination Rate: Message Indicates Disagreement State
Overall	93.8% (405/432)	52.1% (25/48)
Case I	95.5% (190/199)	45.5% (5/11)
Case II	90.0% (99/110)	52.4% (11/21)
Cases III and IV	94.2% (113/120)	53.3% (8/15)

Table 3. Followers' Guesses About the State of the World: by Case.

Case I.

Receive an Agreement State Message			Receive a Disagreement State Message	
Guess that Agreement State	Guess other Agreement State	Guess the Disagreement State	Guess an Agreement State	Guess the Disagreement State
90.2% (359/398)	5.0% (20/398)	4.8% (19/398)	13.6% (3/22)	86.4% (19/22)

Case II.

Receive an Agreement State Message			Receive a Disagreement State Message	
Guess that Agreement State	Guess other Agreement State	Guess the Disagreement State	Guess an Agreement State	Guess the Disagreement State
42.7% (94/220)	0.5% (1/220)	56.8% (125/220)	2.4% (1/42)	97.6% (41/42)

Cases III and IV.

Receive Likelier Agreement State Message			Receive Less-Likely Agreement State Message		
Guess that Agreement State	Guess other Agreement State	Guess the Disagreement State	Guess that Agreement State	Guess other Agreement State	Guess the Disagreement State
86.4% (159/184)	0.0% (0/184)	13.6% (25/184)	44.6% (25/56)	19.6% (11/56)	35.7% (20/56)

Receive Disagreement State Message

Guess an Agreement State	Guess the Disagreement State
3.3% (1/30)	96.7% (29/30)

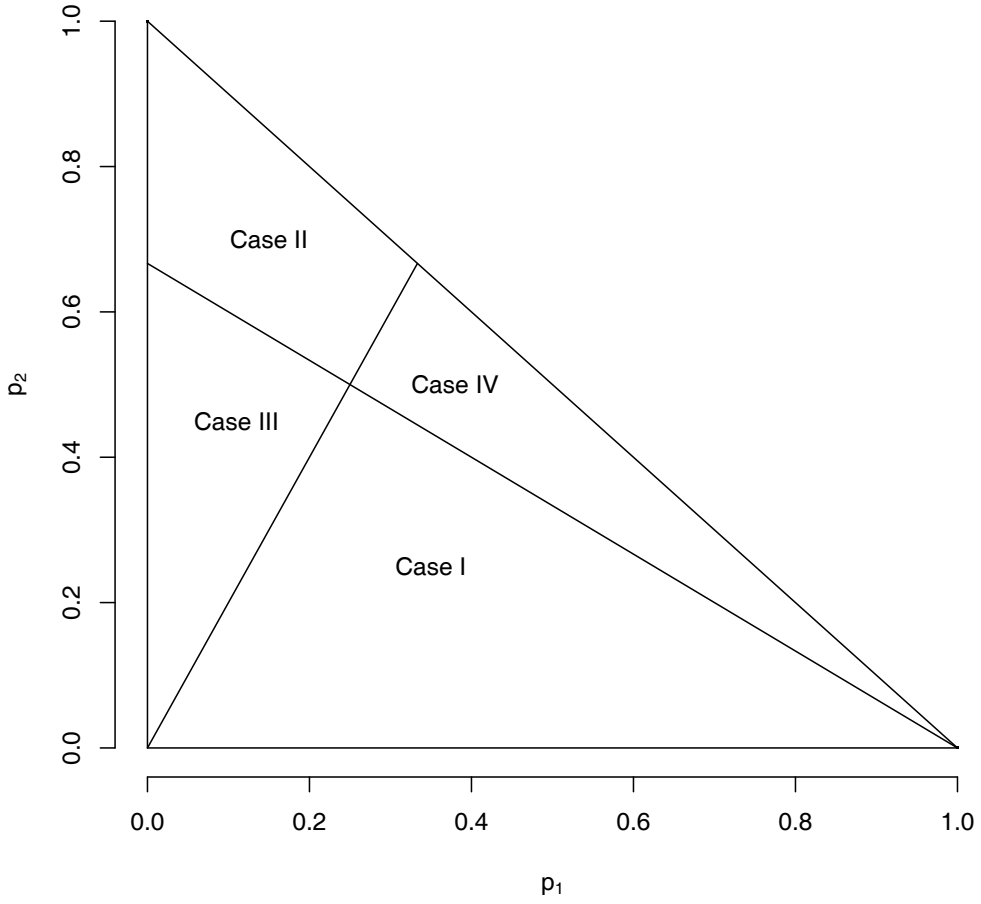

Figure 1: *Graphical Depiction of the Conditions for Cases I-IV on the Probability Simplex.*