

# *Rational Choice Epistemology and Belief Formation in Mass Politics*

Eric S. Dickson\*

Key words: Bounded Rationality, Mass Political Behavior, Belief Evolution,  
Preference Formation, Psychology of Groups

## **Abstract**

This paper begins with a general discussion of the epistemology of rational choice, and argues that there are important questions in political science for which rational choice theory is not a particularly useful epistemic tool. It is further argued that part of the problem lies with the particular vision of methodological individualism that is inherent in the use of classical rational choice assumptions in game theoretic models. An alternative approach that endogenizes the way in which people form beliefs is then advocated as a potential solution to this problem, both as a means to expand the substantive reach of optimizing theories in political science, as well as a way of incorporating more psychological realism into models of political behavior. Two novel models allowing actors within political contexts to form beliefs in endogenous ways are then presented and discussed.

ERIC DICKSON is Assistant Professor in the Department of Politics at New York University. His main research interests include the behavioral foundations of game theory, mass political behavior, and social science experiments. ADDRESS: Department of Politics, New York University, 726 Broadway, Room 744, New York, NY 10003-9580, USA [email: eric.dickson@nyu.edu]

---

\*I wish to thank Jim Alt, Bob Bates, Ethan Bueno de Mesquita, James Fowler, Catherine Hafer, Dimitri Landa, Becky Morton, Ken Scheve, and Ken Shepsle, as well as participants at the Epistemologies of Rational Choice conference, for useful and stimulating conversations related to the ideas contained in the paper. I also thank the anonymous reviewers for their helpful suggestions. Earlier drafts of this paper were under the title "Methodological Individualism, Mass Politics, and Rational Choice Epistemology."

# 1 Rational Choice Epistemology and the Study of Beliefs

When asked to describe the methodology of his field, a nuclear physicist once said, “we learn about the structure of nuclei by smashing them together and seeing what happens.” In this spirit, the study of politics might reasonably be deemed the nuclear physics of the social sciences – politics is much of “what happens” when individuals with divergent interests are thrown together and compelled to make collective decisions about the allocation of power and the distribution of scarce resources in their societies.

But there are, of course, many differences between nuclear physics and rational choice political science; one important difference is an epistemological one, concerning the types of inference typically made in the two fields. The epistemology of experimental nuclear physics generally involves learning about the structure of nuclei based on observations made in the aftermath of nuclear collisions. Such learning takes place by deducing the different macro-level consequences of alternative, and distinguishable, micro-level possibilities, and then observing *which* macro-level consequences are actually realized when the objects of study (nuclei) are exposed to different experimental treatments (smashed together in various combinations and in different ways). Various inferences about the structure of nuclei are then made based on which macro-level consequences have been observed. In short, the fundamental goal is to further understanding of the subject’s microfoundations.

The epistemology of rational choice in positive political theory involves learning of a very different kind. Typically, positive rational choice models seek to explain, or at least to provide a mechanism for or an account of, *macro-level* phenomena.<sup>1</sup> What might be considered the microfoundations of political science – the cognitive pathways through which individual members of society form political judgments, learn about political questions, or come to make political choices – are generally not the objects of interest for rational choice political theorists. Instead, these aspects of human nature are stipulated by assumption, almost always in the form of standard decision theoretic axioms. Investigation of these microfoundational questions is generally left as an exercise for another field – psychology, perhaps, or the behavioral branch of political science – to the extent that rational choice theorists conceptualize it as a task at all. Indeed, substantial

experimental evidence suggesting systematic deviations in practice from most if not all of the fundamental rational choice axioms<sup>2</sup> has met with remarkably little interest from most applied rational choice scholars.

As such, the style of inference that governs the study of nuclear physics – understanding the micro through studying the macro – is quite distinct from that which is common in rational choice political science, where it is taken for granted that we understand the micro well enough for the purposes of learning about the macro. Of course, nuclear physics and politics are very different fields, and it may not be desirable, or even possible, for both of them to conduct their business according to a similar style of inference. At the same time, the distinctive structure of rational choice epistemology has certain consequences that should not be ignored.

One of these involves what, in strict logical terms, can be inferred when the equilibria of a rational choice model, and their comparative statics, correspond to the actual behavior exhibited by actual people in different situations in the real world. It may be the case, as hoped for, that the model employed has accurately captured the preferences, perceptions, and opportunities characterizing actors involved in the political phenomenon in question. But it may also be the case that this is not true; it may be that one or more of these elements of the model has been misspecified, perhaps in a way that represents a fundamental error in the modeler's understanding of the relevant features of the world, but that this error has been “cancelled out” by using the standard rational choice axioms in a situation where these might not hold. A success of applied rational choice modelling is, in essence, a possibility theorem, indicating one potential explanation for the phenomenon of interest. The strength of any inference to be made beyond that depends on a sound empirical understanding of the psychology of decision making in the setting being studied – just the sort of question that rational choice theory sets aside when it stipulates the microfoundations of judgment and decision making by assumption. The extent to which this is a serious problem, a minor problem, or no problem at all of course depends upon the particular application at hand.

The focus of this paper, however, concerns another consequence of the epistemology of rational choice as typically practiced. In short, it will be argued here that by treating the microfounda-

tions of political cognition as exogenously given rather than as objects of study, many important research questions in political science have remained outside the reach of the rational choice world view. Ultimately, the view advocated here will be that this is the case because strict adherence to the rational choice axioms precludes the possibility that people bring different cognitive machinery to bear in different situations – and that the restrictions this implies undermine the optimizing logic on which rational choice theory is itself founded.

In particular, the paper will focus on the study of political beliefs.<sup>3</sup> Understandably, the theory of rational *choice* is most commonly employed in an attempt to predict the actions ultimately taken by actors. But in principle, rational choice models also ought to be able to contribute to the study of political *beliefs*. After all, rational choice models individuals as efficient consumers of information, whose (posterior) beliefs are, at least in part, endogenously determined in rational choice models, based on the information that the actors receive and Bayes' rule.

Suppose that for a given substantive question, the distribution of political beliefs in a population is of some interest; and suppose further that the researcher notices some recurring pattern in such distributions across a range of different situations. An interesting research question would involve developing an account of such patterns in belief distributions.

As an example, consider the following. One finding in the literature on public opinion is that citizens' beliefs on *factual* questions sometimes differ systematically with partisan affiliation. To take a classic (but infamous) example, Democrats and Republicans held strongly different views on the factual question of whether or not Bill Clinton had had an affair with Monica Lewinsky as initial reports of a potential scandal began to emerge (Green, Palmquist, and Schickler (2002)). Examples of this kind, with respect to a range of political issues (some more and some less substantive), pose a potential challenge to the standard rational choice picture of citizens as Bayesian rational agents, who might reasonably differ from one another in terms of their preferences, but who should not differ systematically in terms of their beliefs about factual questions that are ultimately derived from information widely disseminated in the mass media. These public opinion studies are reminiscent of some early work in social psychology, on the different ways in which members of different (social) groups can perceive different realities in the same situation. In one

well known study, for example, Hastorf and Cantril (1954) determined that supporters of two different college football teams formed radically different judgements about the number and importance of fouls committed by their respective teams, with each team's supporters exhibiting a strong tendency to find more fault in the actions of the opposing team. This was the case in spite of the fact that both teams' supporters had seen precisely the same video recording of the game.

Another example of a potentially important research area involving beliefs concerns the perceptions that members of a national, ethnic, or religious group have about members of opposing groups with whom one's own group is engaged in conflict. Understanding the nature and origins of such perceptions would certainly seem relevant to understanding such questions as the origins of conflict, or the optimal strategies a neutral third party who is intent on reconciling the combatants might wish to pursue. As above, a good, and potentially important, research question might involve trying to account for some interesting pattern in this kind of belief about out-group members – for example, exaggerated beliefs about the negative intentions or qualities of individuals from the other group. Indeed, such stigmatization of out-group members is a common phenomenon in social identity theory (Tajfel 1981).

What might a standard rational choice model have to say in addressing such interesting and substantively important patterns of beliefs? Indeed, what might standard rational choice models have to say in general about where it is that given distributions of beliefs come from? Fundamentally, two things can vary in a rational choice formulation: actors' prior beliefs, and the information which actors receive during the course of play. If, in a rational choice model, two actors have divergent posterior beliefs, they must either have had divergent priors, or else access to different information (or both).

There are many interesting stories about public beliefs that might involve this second possibility. People might be selective, either strategically or non-strategically, in their consumption of news, by reading different newspapers or watching different television programs. Or people might learn about politics through complicated social interactions mediated through social networks rather than via the mass media. In either instance, it is easy to see how equally rational actors might draw different conclusions because they have received different information. But in other

settings, explanations of this kind do not seem particularly satisfying. In the context of partisan polarization with respect to beliefs about some new issue or news event (such as the Monica Lewinsky scandal), one might note that different people can react very differently to important news stories that are widely and quite homogeneously reported across different mass media outlets, even controlling for actual level of exposure to the information. In the context of intergroup conflicts that are of long duration and which are covered by news sources to which populations on both sides have free access, it is similarly difficult to imagine that difference in information alone is sufficient to explain the wide gulfs between a combatant's view of the other, and, say, the view that would be held by a neutral observer of the conflict (much less the view that the other has of herself).

In settings where differential exposure to information does not seem a sufficient cause for persistent patterns of divergent public beliefs, where does the rational choice framework lead? Having controlled for one of the two possible sources of divergent posterior beliefs, the weight then falls on differences in prior beliefs as an account for differences in posterior beliefs. In the context of individual responses to a new and small piece of information that has been added to a large past body of experience that is heterogeneous across individuals, this is a natural story. But as the amount of commonly observed information grows to be very large, maintaining a fixed level of polarization in posterior beliefs requires stronger and stronger assumptions about the degree of polarization of prior beliefs. Beyond a certain point, rational choice effectively has nothing to say on the subject aside from stating that actors' prior beliefs are so strong that people simply believe what they believe.

Indeed, consider the implications for the partisan polarization example above. In a situation where differential exposure to information does not seem to plausibly account for the extent of divergence in posterior beliefs, the rational choice framework "explains" the interesting pattern by saying that Democrats and Republicans must have had different prior beliefs – without having anything to say about where these beliefs might have come from. Democrats simply believe one thing, and Republicans another. The situation is all the more incredible when one notes that similar patterns of belief polarization can frequently be observed as clearly in *newly-arisen*

issues (such as, in the example above, public beliefs about what happened in the course of a freshly-broken scandal) as in ancient ones. And the fact that the correlation between beliefs and partisanship in such cases is typically much stronger than the correlation between beliefs and ideology (Green, Palmquist, and Schickler 2002) would seem to reduce the appeal of an explanation based on different prior beliefs being inherited through differing “worldviews” as might be proxied through policy preferences – and perhaps suggests that what is important instead is something to do with the psychology of group membership.

At their core, then, rational choice models seem less inclined to provide helpful insights into certain questions than others, and in particular, seem less likely to teach us useful things about what it is that actors believe than about what it is that actors do. In this context, it becomes more clear why rational choice models are seldom applied in the study of public opinion – they oftentimes simply do not have much to say. It could be argued that this epistemic deficit goes a long way towards explaining the consistency with which the rational choice and behavioral literatures in political science talk past one another completely. Just as psychological research in the absence of game theoretic logic may have little to say about when, for example, different social identities become more relevant in political campaigns and when they do not, because actors’ strategic incentives are not taken into account, standard rational choice approaches have little to say about patterns of public opinion because they do not take relevant psychological insights into account.

It would, of course, be desirable to possess a more unifying framework, which has the potential to simultaneously address variation in interesting patterns of *choice*, addressed by rational choice theory, as well as variation in interesting patterns of *beliefs*, which this section has argued rational choice theory often cannot adequately address. The next section briefly motivates an alternative framework that might possess some promise for addressing the epistemic shortcomings of rational choice models in the study of beliefs, while potentially allowing modelled choice and decision-making behavior to more closely resemble the findings of behavioral economics and the psychology of decision-making. Specifically, it will be argued that processes of belief formation, rather than being taken as exogenous to strategic models of political interactions, should be considered as

being at least partially endogenous to these interactions. The viewpoint taken will be that rational choice faces the difficulties that it does in part because, in its standard application, it misconstrues the proper meaning of methodological individualism in a game-theoretic context.

## **2 Methodological Individualism and Beliefs in Game Theory and Rational Choice**

A cornerstone of rational choice theory is methodological individualism – the notion that the preferences of, beliefs of, and choices made by particular individuals should be taken as primitive elements in the construction of social scientific theories. The alternative to a worldview based on methodological individualism would presumably take the preferences, beliefs, or actions of groups, rather than of individuals, to be primitive concepts.

Methodological individualism seems a natural starting point for social scientific inquiry. From the point of view simply of clarity, methodological individualism has the advantage that the elements it considers to be primitives can be thought of in a way that is both tangible and natural – “objects” which are easily localized (within individual human heads) and which could at least in principle be measured (if not with the technology of today, perhaps with the neuroscience technology of the future). From a theoretical perspective, as well, this starting point seems promising: human beings are creatures whose physical presence, and whose physical structure, are a result of long-run evolutionary selection pressures, and the cognitive structures that endow an individual with decision-making capabilities are coded for in genes contained within that individual. Group-based theories, the presumptive alternative starting point, tend to face crippling problems of vagueness in definition from the very beginning – what, exactly, does it mean to say that a group believes something? Or that it prefers something? Or even that it has done something? And, given that individual beings carry the genes that ultimately code for their biological structures, wouldn't it be necessary to account for why individuals would adhere to any cognitive practice that might be “optimal” for the group? If they can be demonstrated to have an evolutionary incentive to adhere, the theory would no longer be group-based; if they cannot be demonstrated to have such an incentive, the theory would no longer be believable.

The theory of rational choice, of course, is one operationalization of methodological individualism in social science, and it is a natural one. That individuals subject to biological evolutionary pressures would have developed a sense of their own interests and that they would be inclined to pursue them efficiently seems both profound and unobjectionable. In the context of individual decision making, the story behind rational choice, and the connection between this story and the rational choice axioms, are simple and theoretically compelling. An individual must choose an element from some choice set; each available alternative has some consequence (either a deterministic outcome or a lottery over several potential outcomes), that is encoded in the form of a(n expected) payoff the individual is to receive in the event of that consequence. As such, we have the following picture:

**choice set element**  $\longrightarrow$  **payoff**

If the act of choice is to have the potential to be meaningful at all, the individual must have some sort of well-defined preferences over the potential payoffs; this, in a way that is direct and obvious, induces well-defined preferences over the elements in the individual's choice set, because of the direct link between cause (the choice that is made) and effect (the payoff that is received). And it also makes sense that, if an individual has occasion to learn new information about the payoffs that might correspond to certain elements in the choice set, that such information should be incorporated efficiently into the individual's beliefs – that is, that the individual would be best off using Bayes' Rule. There clearly can be no benefit to misinterpretation, or to self-deception, in this context of individual choice. As such, the standard apparatus of rational decision making would seem to be on firm footing in such contexts.

In a more explicitly interactive context, however, the picture is perhaps not so clear; the chain of logic described above does not carry through, at least not in the same way. As before, of course, an individual must choose an element from some choice set. Now, however, the link between the selection of a given alternative and an attendant consequence is less direct; the choices made by others now intervene, and affect what the consequence of a given individual choice might be.

**choice set element *and* others' choices**  $\longrightarrow$  **payoff**

If we were to take the actions of others as given and fixed at some particular values, we would

of course be effectively back in the world of individual choice. In this instance, we could again conceptualize a direct link between the only remaining input variable (an element in the choice set) and the output (the corresponding payoff). But in the play of a game, of course, the actions of others cannot be known *ex ante*, and therefore cannot be taken as given in this way – game theory is interesting only because what might be best for an individual to do depends on what it is exactly that others will ultimately do. Well-defined preferences over the potential payoffs no longer induce well-defined preferences over the elements in a choice set, at least not in the same way, because the direct link between choices and payoffs has been broken.

What, then, can we infer about how individuals should form beliefs in a game theoretic setting? In settings in which it is sensible to think of actors' beliefs as playing no causal role in the choices they ultimately make, it does not much matter. From the point of view of a simple equilibrium concept such as Nash Equilibrium, for example, an individual's beliefs could be altered exogenously without any impact on the set of equilibrium behaviors; if a given action is a best response to other players' actions, it is a best response, period. But if all situations fell into this category, there would never have been any impulse towards modelling belief updating at all; and indeed, from the point of view of more sophisticated equilibrium concepts, like Perfect Bayesian Equilibrium, in which players' actions must be best responses to others' actions *given their own beliefs*, such an argument becomes more problematic. In this case, an exogenous change in an individual's beliefs clearly *could* change the set of equilibrium behaviors, because actors' best response actions given their beliefs could now be different from what they were before. As such, a perturbation to beliefs can directly affect the *choice* behavior that can be sustained in a Perfect Bayesian Equilibrium.

But if this is true, the fundamental game theoretic logic (“what might be best to do depends on what it is exactly that others will ultimately do”) cannot be restricted to actions alone, and the rationale for exporting the rational choice view of belief updating in the form of Bayes' Rule to game-theoretic settings collapses completely. When beliefs play a causal role in determining which choices can be sustained according to a given equilibrium concept, it therefore becomes plausible to suppose that what might be best to *believe* also depends on what it is exactly that

others do – and what it is exactly that others *believe*. And since posterior beliefs are the fruit of a marriage between prior beliefs and new information, one could also suppose that what might be the best way to *learn* (or update beliefs) depends on what it is exactly that others do – and how it is exactly that others are *learning* (or updating beliefs).

This line of reasoning indicates a fundamental tension between two of the foundational impulses of rational choice – actors who are optimizing agents on the one hand, and the use of Bayes’ Rule and the rational choice axioms on the other – in game theoretic contexts. Retaining the view of actors as genuinely optimizing agents may require abandoning Bayesian belief updating – if actors can sometimes secure better equilibrium outcomes by forming their beliefs in some other way. And retaining the view of actors as Bayesian updaters may require acknowledging that such actors’ behavior may be constrained away from being genuinely optimizing.

This paper takes the position that the former approach is inherently more appealing. First, the fundamental motivation for applying game theory to social science was the crucial insight that our understanding of human behavior might be enhanced and organized through the use of optimizing principles. The rational choice axioms and Bayesian updating are simply tools typically used in pursuit of this end. And second, the balance of experimental evidence suggests that the rational choice axioms and Bayesian updating are, in any event, systematically violated by actual human subjects.

What this position implies is that, at least to some extent, the process through which an individual forms beliefs should be considered an endogenous aspect of a model no less than action choice.<sup>4</sup> If Bayes’ Rule is to be employed rather than some alternative updating algorithm – for example one involving some measure of a confirmatory or other type of bias – that this ought to be justified in equilibrium, or at least with empirical evidence, rather than simply by assumption.<sup>5</sup>

It is worth noting that the tension described here – between the basic spirit of game theory and the application of rational choice axioms to individuals in interactive contexts – might provocatively be posed in terms of the proper interpretation of methodological individualism in the social sciences. In a world in which every man is an island, the style of methodological individualism that makes perfect sense in individual choice contexts would be the *only* sort of methodological

individualism. But in a world characterized by group memberships and strategic interactions, perhaps there are other alternatives. In *any* “methodological individualism” it must of course be the case that individuals form their own beliefs, make their own choices, and optimize relative to their own perceptions of their own needs and goals. A game-theoretic framework which treats belief formation behavior as part of individuals’ strategy spaces, while true to the optimizing impulse which motivates game theory, also provides the potential for a certain meeting of the minds between traditional rational choice approaches and other research traditions, such as social psychology and the behavioral branch of political science. There has always been difficulty in discerning what role there could be for social influences on cognition as psychologists understand them in a world where all actors coolly update their beliefs using Bayes’ Rule. The framework under discussion provides one potentially helpful suggestion: things like group membership can affect individual cognition *in equilibrium* because individual belief updating behavior is itself determined in equilibrium. The way in which people will form beliefs will therefore be influenced by the strategic situation in which they find themselves – and therefore, potentially, by things like group membership, whether it be a partisan affiliation, an ethnic tie, or an adherence to one’s alma mater. Allowing belief updating to be treated endogenously allows for a different and potentially richer interpretation of methodological individualism, in which groups might be something aside from an epiphenomenal curiosity, while the individual autonomy and focus that methodological individualism properly insists upon are nonetheless retained. As such, new theoretical perspectives on mass political behavior may be obtained.<sup>6</sup>

Another happy consequence of such a framework is that the types of belief updating and other cognitive behavior that might emerge endogenously could potentially resemble the findings of behavioral economics and the psychology of decision making more closely than do strict rational choice assumptions. That is, by relaxing classical rationality, it may be possible not only to allow actors to optimize in a more inclusive sense than they could while constrained to be Bayesians and the like, but it may also be possible for individual behavior observed as *outputs* of the model to be more descriptively realistic than the rationality assumptions generally assumed to be *inputs* to models. It is also worth noting that an endogenous framework like the one proposed here allows for

the possibility of systematic variation in the way in which people form beliefs in different settings. Might it not be plausible, for example, to suppose that something fundamentally different is going on when individuals form religious beliefs – which can be sustained across generations and over centuries on the basis of nothing that an outside observer would be likely to consider as genuine evidence – as opposed to political ones, which may sway passions, but which are nowhere near as enduring? Or beliefs about the best type of mayonnaise to buy at the grocery store, which may sway no passions at all? This potential to account for variation is what gives the framework a possibility of explaining empirically observed patterns of beliefs, of the kind that initially motivated the paper.

The foregoing discussion has been rather abstract. The remainder of the paper focuses the discussion by presenting two example settings in which one might expect Bayesian belief updating to be a strategically suboptimal cognitive practice – and therefore, to the extent that human agents are genuine optimizing agents, settings in which one might expect Bayesianism to represent a poor approximation to human behavior.<sup>7</sup> Each of these settings describes a different type of strategic benefit that can potentially accrue to actors who deviate from classically rational processes of belief formation. The first model describes a setting in which biased belief formation can assist actors who face problems of *coordination*. The second describes a setting in which biased belief formation can assist actors who face problems of *commitment*. These two strategically distinct settings provide different incentives for belief formation, as well as posing different problems of empirical testability, as will be discussed later.

Before proceeding to the models in detail, however, it is important to make a foundational point about the process by which actors are assumed to arrive at equilibria. Of course, it would seem unnatural to suppose that individuals consciously choose whether to process new information in a biased or in an unbiased way. Instead, the interpretation favored here is that perceptual and information-processing “strategies” are selected for during the course of an evolutionary process. If the process in question is governed by a payoff-monotone dynamic, that is, if the growth rates of strategies’ prevalence in the population is ordered by their expected payoffs against the current population of strategies, then the set of possible outcomes of the process is the same

regardless of the details of the dynamic in question.<sup>8</sup> As payoff-monotonicity is satisfied both by the replicator dynamics model commonly used in evolutionary biology as well as by models of social adaptation such as imitation or reinforcement learning by boundedly-rational actors (Fudenberg and Levine 1998), the models in this paper and their results are applicable regardless of whether the underlying dynamic process is governed by biological or by social evolution. The equilibrium concept that will be employed here, that of evolutionary stability, is intimately related with the familiar concept of Nash Equilibrium in static games – indeed, the evolutionarily stable states of an evolutionary game constitute a subset of the Nash Equilibria of the corresponding static game (Weibull 1995).

The structure of the remainder of the paper is as follows. Section 3 contains a brief introduction to evolutionary game theory that is helpful both in setting the stage for the examples to follow, as well as for motivating the style of modelling advocated here more generally. Section 4 presents the two example models discussed above, illustrating the logic of endogenous departures from orthodox belief formation in coordination and in commitment settings. Section 5 discusses some methodological issues, as well as some questions for future research, and then concludes.

### **3 Evolutionary Games and Evolutionary Stability**

This section contains a very brief introduction to a few essential concepts from evolutionary game theory, and in particular to the equilibrium concept of evolutionary stability that is applied in the paper. Readers who are familiar with this material are encouraged to proceed directly to the next section. Those who desire a more thorough exposition of the subject can find one in recent texts by Weibull (1995) and Vega-Redondo (1996).

Whereas classical game theory is concerned with the behavior of rational actors during a particular play of a game, evolutionary game theory concerns itself with the ways in which members of a population will come to play a game over some period of time. Typically, it is assumed that a large population of agents exists, each of these agents being “programmed” to play some strategy or other in a given game.<sup>9</sup> The individual agents themselves are not the primary objects of interest; they are merely carriers of particular strategies. What *is* taken to be of interest is the

population dynamics of strategies - that is, the way in which particular strategies will come more into, or fade out of, use over time in the population.<sup>10</sup> At any given moment, the distribution of strategies in the population is said to compose a strategic “ecology,” which is characterized by the fraction of the population playing each strategy.

Although different models vary, the most common approach and the one adopted here is that individuals are repeatedly drawn at random from the population in order to play some game with one another. Any changes in the strategic ecology are driven by the payoffs received in plays of the game. Rather than speaking of payoffs to individuals, one speaks of payoffs to strategies, for example, the payoff to strategy  $x$  when facing strategy  $y$  in a game is denoted  $u(x, y)$ . Because the population is large, the average payoff received by strategy  $x$  at some given time can be written as  $\sum_i p_i u(x, i)$ , where  $i$  indexes all strategies in the population (including  $x$ ) and  $p_i$  is the present fraction of the population playing strategy  $i$ . Thus, the average payoff, or *overall fitness*, of strategy  $x$  at any given time is just the population-weighted sum of the payoffs this strategy receives against all other strategies in the population. Crucially, this fitness level depends upon the composition of the strategic ecology.

In a payoff-monotone dynamic, the population fraction of a strategy with higher overall fitness in a given ecology grows relative to the population fraction of a strategy with lower overall fitness in that given ecology. As such, the ecology itself changes over time. Eventually, however, the population might arrive at a state in which all surviving strategies have the same overall fitness; if this eventuality comes to pass, the system is said to be in a stationary state.

The equilibrium concept that is used in this paper, evolutionary stability, involves stationary states that obey a further stability requirement. The idea is as follows. Suppose that a given strategy (pure or mixed) is played by every member of some large population; for this strategy  $x$ , which will be called the “incumbent” strategy,  $p = 1$ . Now inject into this homogeneous population some small population fraction  $\epsilon$  that plays a “mutant” strategy  $y \neq x$ . Now, when individuals from this combined population are randomly paired with one another to play the game in question, they will be drawn against opponents playing the incumbent strategy with probability  $1 - \epsilon$ , but will be drawn against opponents playing the mutant strategy with probability  $\epsilon$ . In

this ecology, the fitness of the incumbent strategy is  $(1 - \epsilon)u(x, x) + \epsilon u(x, y)$ , while the fitness of the mutant strategy is  $(1 - \epsilon)u(y, x) + \epsilon u(y, y)$ . What will happen to the mutant strategy when given this toehold in the population? If  $(1 - \epsilon)u(x, x) + \epsilon u(x, y) > (1 - \epsilon)u(y, x) + \epsilon u(y, y)$  for all  $\epsilon$  below some finite level  $\epsilon^*$  - i.e., if the incumbent strategy  $x$  outperforms the mutant  $y$  for all values of  $\epsilon$  up to some positive threshold - then  $x$  is said to be stable to the introduction of  $y$  so long as  $\epsilon < \epsilon^*$ . That is, any small population playing  $y$  that might come into existence will not endure. If this condition holds for every  $y \neq x$ , then the incumbent strategy  $x$  will be said to be *evolutionarily stable*. Thus, evolutionary stability implies not only that it is a stationary state for a given strategy to dominate a population, but also that this strategy has some degree of robustness against “invasion” as long as the population of invaders is not too large.<sup>11</sup>

An equivalent, and more common, way of presenting evolutionary stability is as follows.  $x$  is evolutionarily stable if and only if (Proposition 2.1 of Weibull 1995)

$$u(x, x) \geq u(y, x) \forall y \neq x$$

and

$$u(x, x) = u(y, x) \implies u(x, y) > u(y, y) \forall y \neq x.$$

In other words,  $x$  must do at least as well against itself as every other strategy does. If this were not the case - if another strategy performed better than  $x$  did against  $x$  - then a small invasion by this alternative strategy would grow, and  $x$  could not be evolutionarily stable. In addition, there is a further requirement: if there are strategies that do just as well against  $x$  as  $x$  does against itself, then  $x$  must outperform all of those other strategies against *themselves*. When the contests against the widespread incumbent strategy are a wash, contests against the invader form a tiebreaker. From the first condition given above, it is clear that evolutionary stability is a refinement of Nash equilibrium.

Thus, evolutionary theory provides a useful alternative interpretation of equilibrium to the familiar one involving calculating, foresighted actors. Individuals are not assumed to make conscious choices to be biased or to be unbiased processors of information; instead, it is assumed that, over time, more successful cognitive strategies outperform less successful ones, and that the behavior we observe represents the equilibrium outcome of such an adaptive process.

## 4 Models of Endogenous Belief Formation: Examples

### 4.1 Endogenous Belief Formation and Coordination

In order for non-Bayesian behavior to be optimal in some setting, it must be the case that the efficient pursuit of “truth” about the world and the efficient pursuit of one’s *interests* are imperfectly aligned. This subsection focuses on one particular set of problems in which it is argued this might be the case: problems involving coordinated or collective action.

The intuition for the argument is as follows. In many situations requiring members of a group to coordinate, the details of their joint plan of action may be less important than that the group coordinate harmoniously on *some* plan of action. Suppose that there exist a number of alternatives on which members of a group could coordinate. Each member of the group has underlying preferences over the different alternatives, but these preferences depend on some feature of the state of the world that is imperfectly known. (For example, an individual might have particular preferences over alternative tax policies if she knew the true state of the economy. However, if the true state of the economy is imperfectly observable, then she cannot know her “true preferences” over tax policies.) As such, individuals in the group must form “perceived preferences” based on their individual beliefs about the state of the world. (Though she does not know the true state of the economy, she has beliefs about the true state of the economy, and these beliefs induce “perceived preferences” over the different tax policy alternatives.) If the perceived preferences of all group members are aligned – for example, if they all perceive a particular alternative to be best – then it will probably be comparatively easy for the group to coordinate on this option. If, on the other hand, individuals have divergent perceived preferences, the attendant coordination problem may be harder to solve. Keep in mind that an individual’s perceived preferences depend upon her beliefs about the state of the world. As such, an individual’s perceived preferences might differ depending upon whether she is a Bayesian or whether she is biased in some manner. If a particular form of bias for an individual can tend to make the coordination problem of the group easier to solve, then it is possible that the individual will be better off with biased perception than she would be if she had formed an efficient view of her own interests.

If this intuition accurately captures a real tension between an individual’s accurate perception

of her own interests and the expected efficacy of group action, then non-Bayesian behavior might be expected to be widespread. After all, many of the ends desirable to human beings are achievable primarily or exclusively through the solution of coordination problems. These include many of the ends that are normally taken to be the concern of political scientists – and most of the ends in subjects such as mass political behavior. For example, elections are often explicitly described as coordination problems that must be overcome by coalitions whose interests may be diverse (Cox 1997). Within the electoral realm, decisive coordination on a candidate or on an ideological program by a political party or faction may be of sufficient value in itself that selection of the very best potential candidate or program may be a less than essential goal. Within the contemporary American system, for example, primary election fights are often seen as costly for a party, since scarce election funds must be depleted, an unattractive and fractious party image can be conveyed, and residual bad feeling within the party can harm mobilization efforts and party cohesion in the subsequent general election. Of course, that costly primary fights are sometimes observed, and that coalitions sometimes suffer coordination failures, demonstrates that biased perception does not serve as a panacea for all ills afflicting group action. At the same time, however, the existence of these phenomena implies neither that biased perception does not exist, nor that it is completely ineffective if it does exist. Much has been written about the ways in which coordination can be achieved since the problem was given its earliest thorough exposition by Schelling (1960); this subsection advocates the view that under some circumstances, coordination problems can be effectively confronted at the level of preference formation by non-Bayesian actors.<sup>12</sup>

Consider the following simple scenario. There are two players, **1** and **2**, each of whom must choose either of two options, **A** or **B**. The interaction between the two players is a coordination game; each player prefers to choose **A** (**B**) given that her opponent will choose **A** (**B**). However, the precise nature of the payoffs is *ex ante* uncertain. Depending upon the state of the world, player **1** might prefer to coordinate with her counterpart on option **A** or on option **B**, and similarly for player **2**. Again depending upon the state of the world, players **1** and **2** might share a common preference as to which option is more desirable, or their preferences on this question might differ. In particular, suppose that there are four states of the world, to be labelled  $\{aa, ab, ba, bb\}$ . For

each of these labels, the first letter designates the preferred option for player **1**, for example  $a$  if player **1** prefers that both players coordinate on option **A** rather than on option **B**, and the second letter designates the preferred option for player **2** in the same way. The set  $\{aa, ab, ba, bb\}$  will be denoted  $w$ , with generic element  $w_j$ , and the probability corresponding to any element  $w_j$  will be represented by  $\rho(w_j)$ .

Further, for each player, the payoff from a coordination failure is normalized to 0, the payoff from coordination on the preferred option is normalized to 1, and the payoff from coordination on the unpreferred option is taken to be  $\mu$ , where  $0 < \mu < 1$  and  $\mu$  is the same for both players. For reasons that will become apparent in a moment, these payoffs will be referred to as the players' *true payoffs*. Each state of the world corresponds to a true payoff matrix as follows:

True Payoff Matrix for State of the World  $aa$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	1, 1	0, 0
Player 1: <b>B</b>	0, 0	$\mu, \mu$

True Payoff Matrix for State of the World  $ab$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	1, $\mu$	0, 0
Player 1: <b>B</b>	0, 0	$\mu, 1$

True Payoff Matrix for State of the World  $ba$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	$\mu, 1$	0, 0
Player 1: <b>B</b>	0, 0	1, $\mu$

True Payoff Matrix for State of the World  $bb$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	$\mu, \mu$	0, 0
Player 1: <b>B</b>	0, 0	1, 1

So far the setup has described the (true) payoffs received by players in different states of the world; it remains to describe the means by which players perceive and process relevant information about the states of the world. In particular, suppose that each player has some initial belief as to which coordination state is better for herself. Then, suppose that each player is sent a fully informative (i.e., noiseless) signal as to which option is in fact a preferable coordination point for herself.<sup>13</sup> In order to explore the possibility that biased perception might be a component of behavior in equilibrium play in this setting, consider the following perceptual “strategy” for a player  $i$  whose initial inclination is to prefer coordination on option **A** (**B**). If she receives a signal

indicating that her prior belief was correct - namely that option **A** (**B**) is a better coordination state for herself - then she correctly perceives this signal with probability 1. However, if she receives a signal indicating that her prior belief was incorrect - namely that option **B** (**A**) is better for herself - then she correctly perceives this signal with probability  $\sigma_i$ , where  $0 \leq \sigma_i \leq 1$ . Players for whom  $\sigma_i < 1$  exhibit a form of confirmatory bias because they sometimes misinterpret evidence against their prior beliefs as being supportive of their preexisting views.<sup>14</sup> Players who are biased in this manner are taken to be unaware of their bias; they assume unquestioningly that they perceive the world as it actually is.

It is assumed that each player's *perception* of her payoffs becomes common knowledge for both players, so that the jointly *perceived* state of the world is a function of both players' perceptual strategies  $\{\sigma_1, \sigma_2\}$ . Depending on the values of  $\sigma_1$  and  $\sigma_2$ , there may be up to four different perceived states of the world, to be labelled  $\{\alpha\alpha, \alpha\beta, \beta\alpha, \beta\beta\}$ . For each of these labels, the first letter designates the option player **1** *perceives* as preferred, for example  $\alpha$  if player **1** perceives it to be in her best interests for both players to coordinate on option **A** rather than on option **B**, and the second letter designates the perceived preferred option for player **2** in the same way. The set  $\{\alpha\alpha, \alpha\beta, \beta\alpha, \beta\beta\}$  will be denoted  $\omega$ , with generic element  $\omega_k$ , and the probability that the perceived state of the world will be  $\omega_k$  given the true state of the world  $w_j$  and players' perceptual "strategies"  $\sigma_1$  and  $\sigma_2$  will be written  $\rho(\omega_k|w_j, \sigma_1, \sigma_2)$ .

After signals about the true state of the world have been received, and converted into perceived states of the world, there are four possible perceived payoff matrices, one corresponding to each of the perceived states of the world:

Perceived Payoff Matrix for Perceived State of the World  $\alpha\alpha$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	1, 1	0, 0
Player 1: <b>B</b>	0, 0	$\mu, \mu$

Perceived Payoff Matrix for Perceived State of the World  $\alpha\beta$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	1, $\mu$	0, 0
Player 1: <b>B</b>	0, 0	$\mu, 1$

Perceived Payoff Matrix for Perceived State of the World  $\beta\alpha$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	$\mu, 1$	$0, 0$
Player 1: <b>B</b>	$0, 0$	$1, \mu$

Perceived Payoff Matrix for Perceived State of the World  $\beta\beta$ .

.	Player 2: <b>A</b>	Player 2: <b>B</b>
Player 1: <b>A</b>	$\mu, \mu$	$0, 0$
Player 1: <b>B</b>	$0, 0$	$1, 1$

Play then proceeds based on the payoff matrix that players perceive. The following assumptions about the outcomes of play in these perceived games are meant to reflect the intuition that certain coordination problems are easier to overcome than others.

If both players perceive a particular option as being best for both of them, both players will be assumed to select this option with probability 1. In other words, players will choose the strategy pair **(A,A)** in perceived state of the world  $\alpha\alpha$ , and they will choose the strategy pair **(B,B)** in perceived state of the world  $\beta\beta$ . This assumption is meant to reflect the fact that coordination in these perceived games is relatively easy because the players' incentives are perfectly aligned, and an obvious focal point exists.<sup>15</sup>

Now suppose that the two players perceive their preferences over the two options to differ. This will be the case in the perceived states of the world  $\alpha\beta$  and  $\beta\alpha$ . Then it is assumed that the players will fail to coordinate with probability  $0 < \mathcal{F} < 1$  - that is, the outcome of play will be either **(A,B)** or **(B,A)**. Alternatively, they may coordinate successfully on **(A,A)** or **(B,B)**, where it is assumed that each of these outcomes transpires with probability  $\frac{1-\mathcal{F}}{2}$  since the perceived game form is symmetric.<sup>16</sup> The introduction of a nonzero probability of coordination failure is meant to reflect that coordination is harder to achieve when incentives are imperfectly aligned than when players think themselves to be in complete agreement.<sup>17</sup>

Because players may or may not perceive the state of the world correctly, the payoff matrix that players perceive may or may not correspond to the true payoff matrix reflecting the actual state of the world. While players make choices based on the payoff matrix that they perceive, their ultimate payoffs are derived from the true payoff matrix corresponding to the actual state of the world. In particular, denote the payoffs received by a player  $i$  who is in the true state of the world  $w_j$ , but who believes she is in the perceived state of the world  $\omega_k$ , as  $\Pi_i(\omega_k|w_j)$ . For example, the value  $\Pi_1(\beta\beta|aa) = \mu$ . This is because players who perceive the state of the world

to be  $\beta\beta$  are assumed, as described above, to play a strategy profile  $(\mathbf{B}, \mathbf{B})$ , and because the true payoff matrix corresponding to the actual state of the world  $aa$  designates the payoffs to be  $(\mu, \mu)$  when these strategies are played.

The *ex ante* payoffs  $U_i$  to each player can be expressed as a sum over all of the possible states of the world and over all of the different ways in which these states can be perceived:

$$U_i(\sigma_1, \sigma_2) = \sum_{w_j \in w} \sum_{\omega_k \in \omega} \rho(w_j) \rho(\omega_k | w_j, \sigma_1, \sigma_2) \Pi_i(\omega_k | w_j). \quad (1)$$

In the usual way, a “strategy”  $\sigma_i$  is a best response to an opponent’s “strategy”  $\sigma_{-i}$  if  $\sigma_i$  maximizes the value of  $U_i$  given  $\sigma_{-i}$ , and a Nash equilibrium in “perceptual strategies” exists when both actors simultaneously choose best responses to one another. Since it seems unnatural to suppose that individuals consciously choose their own cognitive practices, imagine that the coordination game is played by individuals who are repeatedly drawn from a large population, and that “perceptual strategies” that perform well are selected for over time, in the sense related in the previous section.

It remains only to specify the probabilities of different states of the world  $\rho(w_j)$  and the relationship between these and players’ initial beliefs. The following Proposition details the evolutionarily stable perceptual states in the evolutionary coordination game in those circumstances when each player initially believes that the same option is best for herself (eg, the initial condition is that initial beliefs are aligned), and each player’s initial belief is correct with (independent) probability  $q$ .<sup>18</sup>

**Proposition 1. Evolutionarily Stable States when Initial Beliefs are Aligned.** When  $q \leq \frac{1}{1+\mu}$ , all Nash equilibrium outcomes are symmetric and the only evolutionarily stable strategies are

- (1)  $\sigma = 1$  when  $\mathcal{F} \in (0, \frac{1-\mu}{1+\mu}]$ , and
- (2)  $\sigma = 1$  and  $\sigma = 0$  when  $\mathcal{F} \in (\frac{1-\mu}{1+\mu}, 1)$ .

When  $q > \frac{1}{1+\mu}$ , all Nash equilibrium outcomes are symmetric and the only evolutionarily stable strategies are

- (1)  $\sigma = 1$  when  $\mathcal{F} \in (0, \frac{1-\mu}{1+\mu}]$ , and

(2)  $\sigma = 1$  and  $\sigma = 0$  when  $\mathcal{F} \in (\frac{1-\mu}{1+\mu}, \frac{1-\mu}{(1+\mu)(2q-1)})$ , and

(3)  $\sigma = 0$  when  $\mathcal{F} \in [\frac{1-\mu}{(1+\mu)(2q-1)}, 1)$ .

**Proof of Proposition 1.** The proof appears in the Appendix.

The intuition behind these results is as follows. Biased perception can aid in the solution of coordination problems by converting true payoff matrices with “difficult” coordination problems into perceived payoff matrices in which coordination is easier. This maneuver comes at a cost, however – biases in perception will sometimes prevent an individual from knowing her own best interests when these interests might have been attainable. Whether or not biased perception is instrumentally preferable to unbiased perception depends on which loss is worse on average: the inefficiencies entailed in failing to coordinate, or the inefficiencies involved in coordinating consistently, but on unpreferred alternatives.

In the model, this tradeoff is encapsulated in the values of  $\mu$  and  $\mathcal{F}$ . When  $\mu$  is large (close to 1) and  $\mathcal{F}$  is also large (close to 1), then outright coordination failure is a pressing concern, but coordinating on the inefficient choice is not such a big deal. As such, biases in perception can be instrumentally valuable. Conversely, when  $\mu$  is small (close to 0) and  $\mathcal{F}$  is also small (close to 0), then outright coordination failures are not so likely to occur, but coordinating on inefficient choices is quite costly. In these circumstances, biases in perception are more likely to be harmful than to be useful. These tradeoffs can be perceived in the equilibrium results given in Proposition 1. For the smallest values of  $\mathcal{F}$ , only unbiased perception ( $\sigma = 1$ ) is evolutionarily stable; as  $\mathcal{F}$  increases, very biased perception ( $\sigma = 0$ ) also becomes evolutionarily stable; and depending on the value of  $q$ , for still higher values of  $\mathcal{F}$ , unbiased perception may cease to be evolutionarily stable. With a bit more care, the intuitions about  $\mu$  can be discerned in the Proposition as well. As  $\mu$  increases, the endpoints of the different regimes of  $\mathcal{F}$  all shift to smaller values. As such, biased perception becomes evolutionarily stable over a wider range of  $\mathcal{F}$ , and unbiased perception may become evolutionarily stable over a smaller range of  $\mathcal{F}$  (depending upon the value of  $q$ ).

The contents of Proposition 1 can be interpreted in the following way. Individuals in a given population may be confronted with coordination games that have any values of  $\mu$  and  $\mathcal{F}$ ; the Proposition designates what types of behavior will be observed in the equilibrium play of a

given game that has particular values of  $\mu$  and  $\mathcal{F}$ . That is to say, rather than individuals being characterized by a single value of  $\sigma$  that is employed in all coordination games, instead individuals will employ a context-dependent perceptual response  $\sigma(\mu, \mathcal{F})$  that is appropriate to the situation at hand.<sup>19</sup> Thus, despite the fact that all of the equilibria in Proposition 1 are boundary values ( $\sigma = \{0, 1\}$ ), it still might be the case that a given individual sometimes perceives the world accurately but sometimes is biased in equilibrium. This is an important point because any complete picture of perception must allow both for the robust indications of bias in experiments but also for the well-documented view that information about the political world can sometimes affect individuals' beliefs substantially and in the proper direction (Page and Shapiro 1992, Gerber and Green 1997).<sup>20</sup>

One especially interesting, and potentially especially important, region of the  $(\mu, \mathcal{F})$  parameter space is the area around  $\mu = 1$ . In the limit as  $\mu \rightarrow 1$ , the interval of  $\mathcal{F}$  over which unbiased perception is evolutionarily stable shrinks to the point  $\mathcal{F} = 0$ , and biased perception becomes evolutionarily stable for all other values of  $\mathcal{F}$ . If any group coordination focus works as well as any other, and if it is any trouble at all for group members to hash out differences in individual preference, then it is perfectly sensible for group members to see the world through biased eyes. Indeed, it would be counterproductive were they to do otherwise.

One reason that this parameter region may be of particular importance to the study of mass political behavior lies in the study of political ideologies. As Bawn (1999) writes, "ideology matters in politics because it causes people to care about issues in which they have no direct stake." She continues by characterizing ideology as "creating preference in the absence of interest." Such a setting would of course correspond to a value of  $\mu = 1$ , because in this case individuals would be indifferent between the different coordination alternatives. The results of the model suggest that, in relation to ideological matters as Bawn defines them, individuals should be especially prone to interpret the world in a biased rather than in an unbiased way.

The parameter region near  $\mu = 1$  is also important because it has implications for the distributions of public opinion that should be expected to be observed in practice. Consider the coordination problem faced by a diverse group of individuals who must select a plan of action

encompassing a number of policy areas that do not logically constrain one another. Suppose that the membership of the group shares a common interest with respect to one particular policy area that is considered important by everyone (say, economic policy), but interests differ on a number of other policy areas, each of which is by itself considered relatively unimportant compared to the policy area on which there is general agreement. As in the model above, individuals are taken to have true preferences over each of these policy areas, but these true preferences depend upon information about the state of the world that is imperfectly known. As such, an individual's perceived preferences will be influenced by the way in which she processes information about the state of the world. What will be the optimal way in which to update beliefs that are relevant to one of the *less* significant policy areas?

Consider a scenario in which actors receive one signal that is relevant to their views on only one of these minor issues. The optimal way of processing this signal will be determined according to the model by the tradeoff between the pursuit of individually-preferred outcomes and the effectiveness of group coordination. Because there are multiple issues, the outcomes in question in the coordination game are bundles including all of the multiple issues. However, the information being received has a bearing on one minor issue only; as such, the relevant value of  $\mu$  must be large (near 1), because the one minor issue in question is relatively unimportant to individuals' views on the outcomes as a whole. But this means, with respect to information processing about the minor issue, that biased perception is likely to be in equilibrium.<sup>21</sup> Because the same logic holds for *all* of the minor issues, a likely consequence is that members of the group can be expected to maintain correlated preferences across different policy areas. This is true even though these different policy areas are logically unconnected. This implication resonates with natural intuitions about ideology and ideologues: a gay rights supporter is more likely to favor the same sort of environmental policy as another gay rights supporter than she is to favor the same sort of environmental policy as a gay rights opponent, even though gay rights and environmental policy basically have nothing to do with one another.<sup>22</sup> As such, understanding the potential role of biased perception in solving coordination problems may also aid in understanding the distributions of political preferences in mass publics.

## 4.2 Endogenous Belief Formation and Commitment

One of the central problems of conflict resolution is the reconciliation of actors whose interests are at odds, and who may reasonably doubt each others' motives in the early stages of settlement negotiations. Such a process of reconciliation is especially difficult when the conflict it is intended to resolve has been intense and long-lasting. Factions that have clashed in civil war may be reluctant to expose themselves to trickery by dangerous opponents. A government may doubt the intentions of a terrorist organization that claims it wishes to lay aside arms and negotiate peacefully – while the terrorist organization may not trust the government to carry out its side of the bargain once disarmament has taken place. At bottom, the problem is one of building, and maintaining, trust between actors.

One of the most fruitful approaches to modelling such peacemaking processes is the reassurance framework laid out by Kydd (2000), in which actors make concessions to their counterparts which can, in some circumstances, send an informative signal about the trustworthiness of the conceiver. Like most formal models of intergroup conflict, the reassurance framework assumes that actors share common knowledge of each others' (Bayesian) rationality as well as common priors about the structure of the uncertainty inherent in the situation. And yet, a considerable literature within international relations theory emphasizes the doubts actors may have about the rationality of their opponents, as well as the incentives actors may sometimes have to cultivate such doubts in others (the "rationality of irrationality"). Indeed, it is readily apparent that adversaries in some conflicts display beliefs about the "other" that can seem distorted, extreme, or even bizarre to neutral (but historically-informed) outside observers. As such, it seems worth considering relaxations of standard assumptions of common knowledge of rationality in the emotionally charged settings common to problems of conflict resolution.

Indeed, there is good reason to believe that actors with certain types of biases may well hold a *strategic advantage* over Bayesian actors in settings such as reassurance games. Suppose a given actor is somewhat more suspicious of her counterpart than Bayes' Rule and the objective information at hand would necessarily allow. And imagine further that this excessive degree of suspicion is at least partially observable. How might this affect the counterpart's behavior?

If the counterpart is trustworthy – in the sense that he would genuinely prefer a reasonable peace agreement to continued conflict, and would not wish to take advantage by renegeing on his commitments partway through the peace process to resume conflict on more favorable terms to himself – then he might be willing to indulge the doubter’s excessive suspicion by offering some greater degree of concessions, at least up to a point. Thus, an excessive degree of suspicion may yield material benefits. But, perhaps more importantly, there may be an *informational* advantage as well. Suppose that the level of excessive suspicion, and the degree of concessions that would be required to allay this suspicion, are large enough that a trustworthy counterpart *would be* willing to indulge the doubter and bear the concession costs while an untrustworthy counterpart *would not*. If this is the case, being excessively suspicious may not only elicit direct benefits in the form of greater concessions, but may also allow an actor to *learn more* about the nature of her counterpart based on whether or not she is willing to offer such concessions.

The model presented in this section allows the way in which one actor forms beliefs about another to emerge endogenously; the argument will be that, by allowing an actor to form non-Bayesian posterior beliefs in a way determined by the specific nature of the strategic interaction, it may be possible for her to achieve better outcomes because such non-Bayesian posterior beliefs may allow the actor to overcome a particular kind of commitment problem. An actor commonly known to be Bayesian rational might *wish* to claim that she would only engage in a risky peace process if some higher level of concessions were offered – but such a claim would not be credible. An actor possessing in-equilibrium non-Bayesian beliefs dictating an excessive degree of suspicion *could* be credible in making such a claim, and therefore could potentially achieve better outcomes.<sup>23</sup>

#### 4.2.1 Basic Reassurance Model

Consider a traditional reassurance model in which there are two actors: Eve and Adam.<sup>24</sup> Adam can be of either of two types: “good” or “bad”; his type, drawn from a commonly known binomial probability distribution, is “good” with probability  $p_0$  and “bad” with complementary probability  $1 - p_0$ . Adam knows his own type, but Eve knows only the probability distribution. There are two stages to the reassurance game; Adam has the opportunity to choose a level of concessions that may (or may not) affect Eve’s in-equilibrium beliefs, and then the two players engage in a

trust game, the outcome of which ultimately depends upon Adam's type and Eve's belief about Adam's type. All these aspects of the game structure are common knowledge.

Specifically, in the first stage, once Adam's type has been drawn, he must choose a level of concessions  $x \in \{0, \tilde{x}\}$ , where  $x = 0$  represents no concession and  $x = \tilde{x} \in (0, 1)$  represents a positive concession. The choice of concession is not cheap talk, but rather affects payoffs in two ways. First, the concession results in a direct transfer from Adam to Eve; a concession  $x$  results in an addition of  $C(x)$  to Eve's utility, and a subtraction of  $C(x)$  from Adam's. We take  $C(x = \tilde{x}) = c > 0$  and  $C(x = 0) = 0$ . Second, the choice of  $x$  is also taken to affect the *probability* of victory that each side enjoys if there is conflict, in a way to be described shortly.

In the second stage, Eve must decide whether to take an action  $C$  that extends the potential for cooperation (and a peaceful settlement), but also allows for the possibility of exploitation if Adam is of the bad type, or an action  $D$  that represents an uncooperative move precluding a peaceful settlement. If Eve chooses  $D$ , the reassurance game therefore ends, and conflict ensues; but if Eve chooses  $C$ , then Adam in turn has the opportunity to choose to cooperate ( $C$ ) or not to ( $D$ ). If Adam also chooses the cooperative action,  $C$ , then a peaceful settlement is achieved; but if he chooses the uncooperative action,  $D$ , then conflict results on terms more favorable to Adam than there would have been had Eve declined to make a cooperative gesture. Conflict is modelled as a stochastic process that either ends favorably for Adam or ends favorably for Eve.

The detailed structure of the payoffs is as follows. A peaceful outcome yields a payoff that is normalized to 1 for any actor (for Eve as well as for both Adam types), and a conflict ending unfavorably yields a payoff that is normalized to 0 for any actor (for Eve as well as for both Adam types). However, different actors have different relative preferences over outcomes. Specifically, the payoff for a favorable outcome will be  $\beta_{Eve}$  for Eve;  $\beta_{Adam}^g$  for a good type of Adam; and  $\beta_{Adam}^b$  for a bad type of Adam. The types are distinguished behaviorally by assuming that  $\beta_{Adam}^g$  is sufficiently large that a good Adam type would always prefer to return a conciliatory gesture by Eve, while  $\beta_{Adam}^b$  is sufficiently small that a bad Adam type would always prefer to betray a conciliatory gesture by Eve.

If a conflict results after Eve fails to extend cooperation (eg, she plays  $D$ ), then Adam (of

either type) wins the conflict with probability  $x_0 - x$  while Eve wins it with complementary probability  $1 - x_0 + x$ .  $x_0 \in [\tilde{x}, 1]$  can be thought of as a crude measure of Adam's strength relative to Eve's, while  $x$  is the level of Adam's concession in the first stage of the game. Thus, as foreshadowed earlier, a concession made by Adam is potentially costly not only in a direct sense but also in terms of the probability of victory in a potential conflict. If instead a conflict ensues after Adam fails to cooperate (eg, he plays  $D$ ), snubbing Eve's cooperative gesture, then Adam (of either type) wins the conflict with increased probability  $x_0 - x + \epsilon$ , where  $\epsilon > 0$ , while Eve wins it with decreased probability  $1 - x_0 + x - \epsilon$ .

As is typical for signalling games, the game described here has a variety of equilibria.

**Proposition 2. On Equilibria of the Basic Reassurance Model.** There are three different kinds of Nash equilibrium outcome in pure strategies:

(1) There is a separating equilibrium in which the good type of Adam makes a concession  $x = \tilde{x}$  while the bad type of Adam makes a concession  $x = 0$ , so long as  $(\epsilon - \tilde{x})\beta_{Adam}^b \leq c \leq 1 - x_0\beta_{Adam}^g$ , independent of the value of  $p_0$ .

(2) There is a pooling equilibrium in which both types of Adam make a concession  $x = \tilde{x}$ , so long as  $p_0 \geq p^*(\tilde{x})$  and  $c \leq \min\{1 - x_0\beta_{Adam}^g, (\epsilon - \tilde{x})\beta_{Adam}^b\}$ .

(3) There is a pooling equilibrium in which both types of Adam make a concession  $x = 0$ , so long as (i)  $p_0 \geq p^*(0)$  or (ii)  $p_0 < p^*(0)$  and  $c \geq \max\{1 - x_0\beta_{Adam}^g, (\epsilon - \tilde{x})\beta_{Adam}^b\}$ .<sup>25</sup>

**Proof of Proposition 2.** The proof appears in the Appendix.

Thus, the basic reassurance model has both separating and pooling equilibria. All things equal, either type of Adam would prefer that Eve believe he is likely to be of the good type; a good Adam would prefer this because it extends the possibility of a peaceful outcome, while the bad Adam would prefer this because it extends the possibility of exploiting Eve. In the separating equilibria, the good Adam type is able to send a costly signal – through positive concessions – that the bad Adam type is unwilling to match. In the pooling equilibrium on  $x = \tilde{x}$ , concessions are inexpensive enough that the bad Adam type is willing to make them in order to avoid revealing his nature. In the pooling equilibrium on  $x = 0$ , concessions are expensive enough that neither Adam type is willing to make them – or Eve is so convinced that Adam is of the good type that

neither player need make them.

A careful examination of the equilibrium conditions indicates that there are some situations (values of the parameters) in which one type of equilibrium is unique – for example, pooling on  $x = 0$  – and there are others in which there are multiple equilibria – for example, either pooling on  $x = 0$  or separating.

It is important to note that the equilibria described here have different consequences for Eve’s welfare. For example, a pooling equilibrium on  $x = 0$  is somewhat unpleasant for Eve for two separate reasons: she receives no concessions, and in addition, she learns nothing about Adam’s type from observing his actions. On the other hand, a separating equilibrium would appear to be somewhat more appealing; one of the two types of Adam does make a concession, and she does learn Adam’s type. A pooling equilibrium on  $x = \tilde{x}$  offers different benefits relative to the  $x = 0$  pooling equilibrium; both Adam types make concessions, but she does not learn Adam’s type.

The following section allows Eve to deviate from in-equilibrium Bayesian updating about Adam’s type by imagining that the way she updates her beliefs emerges endogenously in the context of an evolutionary game. The question of interest is: can Eve achieve better outcomes if she updates her beliefs about Adam according to a different procedure than that implied in standard treatments of Perfect Bayesian Equilibrium?

#### 4.2.2 Reassurance Model with Endogenously Determined Suspicion

Now suppose instead that Eve forms her beliefs in a different way that reflects a different logic than the familiar “within-equilibrium” Bayesian updating in Perfect Bayesian Equilibria. Specifically, consider the following alternative updating procedure. After nature has selected a type for Adam, but before anything else happens, Eve “adopts a posture” towards Adam. This “posture” describes how Eve’s beliefs would be affected by observing each of the different concessions that are available to Adam. Recall that Adam chooses some level of concession  $x \in \{0, \tilde{x}\}$ . As such, a posture for Eve will take the form  $(\bar{p}(0), \bar{p}(\tilde{x}))$ , where  $\bar{p}(0)$  is the posterior belief she would have that Adam is of the good type if she were to observe a concession of 0, while  $\bar{p}(\tilde{x})$  is the posterior belief she would have that Adam is of the good type if she were to observe a concession of  $\tilde{x}$ . No *a priori* assumptions are made about the contents of Eve’s posture in terms of the specific

probability assessments that they prescribe.

Ultimately, an individual’s “posture” is envisioned as a type-like characteristic that is not consciously chosen. Rather, certain belief-updating types will achieve better outcomes than others do in a population consisting of different types; a posture will be deemed an “equilibrium posture” if it can exist in the equilibrium of an evolutionary game.<sup>26</sup> For purposes of comparison, the term “Bayesian posture” will be employed to refer to a posture which commits Eve to possessing the same posterior beliefs that could be held by a Bayesian agent who observed Adam’s choices in a given equilibrium.<sup>27</sup> For example, in the context of an equilibrium involving Adam types that play separating strategies  $(x_g, x_b) = (\tilde{x}, 0)$  in the concessions they offer, a Bayesian “posture” would specify  $\bar{p}(\tilde{x}) = 1$  and  $\bar{p}(0) = 0$ . In the context of another equilibrium involving Adam types that play pooling strategies where  $x_g = x_b$ , a Bayesian “posture” would specify  $\bar{p}(x_g) = \bar{p}(x_b) = p_0$ . If, in a given equilibrium of the evolutionary game with endogenously determined suspicion, the recipient has an equilibrium (optimal) posture that differs from any Bayesian posture in the analogous setting of the basic reassurance model, this equilibrium posture will be referred to as *biased*.

The sequence of events is envisioned as follows: Nature selects a type for Adam; Eve adopts a posture; Adam decides what concession to make to Eve; Eve, possessing a posterior belief that is determined by the posture she adopted earlier and the concession she observed, chooses whether to play *C* or *D*; and then Adam faces a similar choice if Eve has chosen the cooperative action *C*.

The following proposition demonstrates that in the equilibria of this reassurance model with endogenously determined suspicion, there exist situations in which the postures that emerge in evolutionary equilibrium are in fact biased – that is, Eve can do better by deviating from the standard picture of belief updating within the context of a Perfect Bayesian Equilibrium if she is allowed to adopt a “posture” that simply maps the action she observes the other player to take onto a posterior belief.

**Proposition 3. On Equilibria of the Reassurance Model with Endogenously Determined Suspicion.** Take the equilibria of the (evolutionary) reassurance model with endogenously determined suspicion, and compare them to the corresponding equilibria of the basic reassurance

model.<sup>28</sup> Then:

(1) There exist situations in which the Adam types have a unique equilibrium profile of concessions in the basic reassurance model, but have a *different* unique equilibrium profile of concessions in the evolutionary model, and Eve only adopts biased postures in equilibrium (which leave her strictly better off)

(2) There exist situations in which the Adam types have multiple equilibrium profiles of concessions in the basic reassurance model, but have a unique equilibrium profile of concessions in the evolutionary model, and Eve's equilibrium postures are a subset of the Bayesian postures (which leave her weakly better off than she was in the equilibria of the basic reassurance game).

**Proof of Proposition 3.** The proof appears in the Appendix.

The proposition demonstrates that, in equilibria of the game with endogenously determined suspicion, Eve sometimes adopts biased postures in equilibrium, and that allowing her to update her beliefs about Adam in an endogenous way can leave her strictly better off.

A few examples can highlight some of the intuition behind the result. As indicated previously, the different equilibria of the basic reassurance game have different consequences for Eve's welfare. Consider, as an example, the set of situations in which there are two types of equilibrium: pooling on  $x = 0$  and separation by Adam. Eve would have a clear preference for separation, from a welfare standpoint, because this equilibrium offers a higher level of concessions and of informational benefits than does the pooling equilibrium. Part (2) of the proposition indicates that, in situations such as this one, Eve can effectively *guarantee* that she will end up in the separating equilibrium by adopting a particular posture. Specifically, by adopting a posture that she will trust Adam if he makes a concession but will not trust him if he does not, it turns out to be the case here that she can *enforce* separation; if she were an ordinary Bayesian actor, and if both Adam types decided they wished to offer  $x = 0$ , she would be instead be stuck in an inferior equilibrium. In essence, by forming her perceptions of Adam's intentions in a biased way, she overcomes a commitment problem that would leave an ordinary Bayesian actor with a poorer outcome. In this instance, therefore, Eve's adoption of a posture can be thought of as a sort of equilibrium selection device that helps settle matters on terms more favorable to Eve.<sup>29</sup>

Part (1) of the proposition makes an even stronger claim – there are situations in which Eve can induce completely different patterns of concessions from Adam than would be possible in equilibria of the basic reassurance game. For example, the proof of the proposition demonstrates that there are some situations in which Adam’s types pooling on  $x = 0$  is his unique equilibrium behavior in the basic reassurance game, but in which Eve is able to induce pooling on  $x = \tilde{x}$  instead by adopting a biased posture. As before, the relevant posture is that she will trust Adam if he makes a concession but will not trust him if he does not. In the settings in question, this posture involves Eve requiring a greater show of goodwill from Adam than she would need if she were Bayesian rational. That is, there are certain situations in which the costs of concessions are not too high, but neither Adam type finds it worth his while to make a positive concession because Eve is inclined to trust him anyway. By being excessively suspicious of Adam in some such circumstances, she can induce *both* of Adam’s types to make the concession, yielding a direct benefit to her (though she does not gain any informational benefits). Part (1) of the proposition therefore demonstrates that there are some situations in which Eve only adopts biased postures in equilibrium, and can induce completely different behaviors from Adam than would otherwise have been possible.

## 5 Discussion and Conclusion

The standard rational choice framework has been so useful in addressing such a wide range of phenomena that reluctance to deviate from it is understandable. This is especially the case because our habituated identification of “optimal” behavior with rational choice is so deeply engrained that suggesting an alternative optimizing theory of behavior and belief formation at first glance seems not so much heretical as incoherent. Years of research in psychology and experimental economics have provided ample cause for *empirical* doubts about each of the various planks of the rational choice world view; this paper has presented several *theoretical* justifications for *expecting* optimizing agents to exhibit “behavioral” rather than classically rational belief-updating behavior. In the process, the paper has proposed an alternative framework of analysis, in which belief formation behavior is allowed to emerge endogenously in equilibrium rather than being

assumed to be Bayesian (or anything else); argued that this framework may allow for models that are not only more descriptively realistic than rational choice ones, but which can gain more epistemological leverage as well; and presented a few example models to illustrate some intuitions as to why endogenous models of belief formation might be useful tools of analysis for substantive political questions, as well as to illustrate some possible operationalizations of the advocated method.

Of course, many important concerns and questions stand between these observations and the implementation of a fruitful program of applied research in political science.

One likely objection to behavioral models of this kind focuses on the potential loss that agents might suffer in obtaining biased posterior beliefs about the world. At the end of the day, players receive payoffs based on their actions (and the actions of others), not based on their (or anyone else's) beliefs. Given that this is the case, wouldn't it be better for an agent to "act" like a behavioral agent, but privately to maintain "correct" (that is, Bayesian) posterior beliefs?

The different example models presented in the previous section suggest different potential replies. In the first example model, which involved a coordination setting, the nature – and difficulty – of the challenge faced by players depended on the posterior beliefs they formed. As such, the relevant reply to the critic of behavioral models would be that a perfectly rational agent potentially *could* make the same ultimate strategy choices as a behavioral agent – but that doing so in a properly coordinated manner could potentially be more difficult for perfectly rational agents than it would be for "behavioral" agents. After all, if coordination games with divergent interests could be assumed always to have happy endings, the phrase "coordination *problem*" would never have been coined.

That said, it obviously is true that actors who share a common group membership but who have divergent interests sometimes do manage to achieve effective action while acknowledging their differences. It is an open question – and an empirical question – to what extent coordination problems of the kinds interesting to political scientists are typically solved via "strategic" means rather than at the level of belief formation. The answer, as with most things, probably depends on the context. The "strategic" solution is probably more easily achievable by political elites who are

few in number and who do this sort of thing for a living; members of mass publics, or of partisan or other social groups within mass publics, are perhaps less likely to achieve their goals through conscious efforts at strategic coordination, and more likely to do so effectively by subscribing to an ideology or a group myth that acts as an effective coordination device.<sup>30</sup> Even when coordination is successfully achieved, and the behavioral and strategic stories seem observationally equivalent in terms of *actions*, it may well be possible to distinguish them empirically through other means – for example, through the use of survey research as one means of attempting to measure beliefs directly. Here the two competing approaches can have highly distinct implications even when they do not at the level of choice behavior. The discussion of empirical patterns of beliefs that motivated the paper suggests that some substantial proportion of the evidence may well come down on the side of effective coordination through a mechanism of “behavioral” belief formation.

In the second example model, the reply to the critic of behavioral models would be different. In this model, which involved a commitment problem, the perfectly rational agent is at a loss compared to his behavioral cousin because the behavioral agent can compel different behavior from her *opponent* than the perfectly rational agent can. The behavioral agent can threaten to behave in a way that would not be credible if it came from the lips of the perfectly rational agent. Here the potential difficulty lies in the observability of the agent’s type. As in the models of endogenous preference formation cited earlier, agents’ types are assumed to be observable here. In a world of incomplete information, there would surely be some circumstances in which perfectly rational agents would have an incentive to – and sometimes might be able to get away with – masquerading as behavioral agents. But, in the end, the behavioral agents are different here *because they would be willing to take some actions that the perfectly rational agents would not*. And as such, it is clear that there will be some situations in which the perfectly rational agents will not be able to stomach what would be required to pull off their masquerade successfully.

It is also worth pointing out that the assumptions about the observability of type that are present in the belief-formation models discussed here are in a sense very similar to the implicit assumptions made about the common knowledge of rationality in virtually every applied rational choice model in political science. After all, common knowledge of rationality boils down to (some-

thing more than) transparency of actors' belief-updating mechanisms to other actors. Indeed, in the context of an (orthodox) equilibrium concept like Perfect Bayesian Equilibrium – in which players are taken to share common conjectures about everything from each others' rationality and belief-updating behavior through the actions that a hypothetical type *would* take if it existed – the assumptions about type observability that are made here seem almost modest by comparison. And, it bears repeating that the deviations from Bayesian rationality predicted in the framework presented here are *automatically* credible in the sense that they are determined in equilibrium, and in the sense that actors who gain experience of others would come to observe such behavior to be stable and would come to expect it.

This paper has argued that there are solid theoretical reasons to question the primacy of standard rational choice assumptions because, in a variety of settings, behavioral actors of different kinds might be expected to outperform classically rational ones. But it is very unclear what the correct way to think about the underlying evolutionary process might be – and this is an important question, because the answer may go far toward determining exactly how much flexibility individuals' cognitive structures have in conforming to different strategic situations in practice.

For example, one's expectations about the correct way of modelling human behavior might differ depending upon whether the basic decision-making and belief-formation machinery were fixed and encoded biologically, or whether they emerge through processes of social learning in response to stimuli and rewards in the social environment. More specifically, without an understanding of the microfoundations of the relevant adaptive process, it is unclear from a theoretical standpoint how finely individual cognition patterns could be expected to vary according to the strategic environment. Indeed, this is an open question from an empirical point of view as well; while experimental literatures in psychology and economics provide many grounds for doubt about the standard depiction of human rationality, little experimental work has systematically explored the question of potential variation in the strength or nature of biases as a function of strategic aspects of the environment. At one extreme on the spectrum of possibilities is a scenario in which individuals simply evolve some overall level of, say, bias in processing information that is stable given the constrained optimization problem of carrying the same bias into every situation.

That is, cognitive behavior could be basically exogenous to specific situations, but in a way that might deviate from that generally assumed in models of classical rational choice. If this were the case, modelling cognitive practices as being endogenously determined in specific games would be the wrong approach – unless success in the specific game in question were the dominant factor relevant to the underlying adaptive process. At the other extreme, the level of bias would be endogenously determined in every unique situation an individual confronts. In this case, the level of bias observed in different settings could be seen to trace the comparative static predictions that would result from models of endogenous cognition. It is an open question, and one which could be fruitfully addressed in future experimental work, which pole is a more accurate depiction of human nature (if either is).<sup>31</sup> Individual behavior is in any case a complicated interaction between consciously and unconsciously chosen actions and cognitive practices, and it seems likely that the answer lies somewhere in between.

In the end, models that organically mix evolutionary game theory (to model the cognitive process) and explicit strategic thinking (given the cognitive process, how people choose to behave) are likely to help in the pursuit of the answers. This is likely true not only within individuals, but also across individuals. One of the most striking facts about human cognition revealed in experimental and empirical studies is the large degree of interpersonal heterogeneity in such qualities as the extent of strategic reasoning across laboratory subjects (Camerer 2003). This observation, combined with the rich tradition of behavioral research on varying levels of political sophistication in mass publics, suggests that the modelling of interactions between differently-able (or differently-aware) individuals is likely to be an important subject in future political science research, and may perhaps ultimately assist in our understanding of theoretically elusive topics such as the nature of political leadership.

The answers to such questions – and the potential for our access to such answers – have profound implications for what models of individual choice and belief formation might ultimately be able to teach us not only about politics, but also about ourselves. One day, political science may resemble nuclear physics in ways that today we can scarcely imagine.

## 6 Appendix: Proofs

**Proof of Proposition 1.** Because the labels **A** and **B** are arbitrary, it can be taken without loss of generality that each of the players shares an initial belief that option **A** is better for herself. The following table contains every permutation of  $\omega_k$  and  $w_j$  along with the probability  $\rho(\omega_k|w_j, \sigma_1, \sigma_2)$  and any values of  $\Pi_1$  and  $\Pi_2$  for which the corresponding  $\rho$  is nonzero:

$\omega_k$	$w_j$	$\rho(\omega_k w_j, \sigma_1, \sigma_2)$	$\Pi_1(\omega_k w_j)$	$\Pi_2(\omega_k w_j)$
$\alpha \alpha$	aa	1	1	1
$\alpha \beta$	aa	0	-	-
$\beta \alpha$	aa	0	-	-
$\beta \beta$	aa	0	-	-
$\alpha \alpha$	ab	$1 - \sigma_2$	1	$\mu$
$\alpha \beta$	ab	$\sigma_2$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$
$\beta \alpha$	ab	0	-	-
$\beta \beta$	ab	0	-	-
$\alpha \alpha$	ba	$1 - \sigma_1$	$\mu$	1
$\alpha \beta$	ba	0	-	-
$\beta \alpha$	ba	$\sigma_1$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$
$\beta \beta$	ba	0	-	-
$\alpha \alpha$	bb	$(1 - \sigma_1)(1 - \sigma_2)$	$\mu$	$\mu$
$\alpha \beta$	bb	$(1 - \sigma_1)\sigma_2$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$
$\beta \alpha$	bb	$\sigma_1(1 - \sigma_2)$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$	$\frac{1-\mathcal{F}}{2}(1 + \mu)$
$\beta \beta$	bb	$\sigma_1\sigma_2$	1	1

As a result, the expression for  $U_1$  takes on the following form (the expression for  $U_2$  is analogous):

$$\begin{aligned}
U_1(\sigma_1, \sigma_2) = & q^2[1] + q(1 - q)[(1 - \sigma_2) + \sigma_2 \frac{1 - \mathcal{F}}{2}(1 + \mu)] + \\
& + q(1 - q)[(1 - \sigma_1)\mu + \sigma_1 \frac{1 - \mathcal{F}}{2}(1 + \mu)] + \\
& + (1 - q)^2[(1 - \sigma_1)(1 - \sigma_2)\mu + (1 - \sigma_1)\sigma_2 + \\
& + \frac{1 - \mathcal{F}}{2}(1 + \mu) + \sigma_1(1 - \sigma_2) \frac{1 - \mathcal{F}}{2}(1 + \mu) + \sigma_1\sigma_2] \tag{2}
\end{aligned}$$

Equation [2] can be rewritten as:

$$U_1(\sigma_1, \sigma_2) = \sigma_1 \{ \sigma_2 [(\mu + 1)(1 - q)^2 \mathcal{F}] + (1 - q) [ \frac{1 - \mathcal{F}}{2}(1 + \mu) - \mu ] \} + R_1 \tag{3}$$

while the analogous form for  $U_2$  is:

$$U_2(\sigma_1, \sigma_2) = \sigma_2 \{ \sigma_1 [(\mu + 1)(1 - q)^2 \mathcal{F}] + (1 - q) \left[ \frac{1 - \mathcal{F}}{2} (1 + \mu) - \mu \right] \} + R_2 \quad (4)$$

where the term  $R_1$  ( $R_2$ ) in equation [3] ([4]) is independent of  $\sigma_1$  ( $\sigma_2$ ). Note that both of the equations [3] and [4] are of the form:

$$U_i(\sigma_i, \sigma_{-i}) = \sigma_i \{ A\sigma_{-i} + B \} + R_i \quad (5)$$

where  $A = (\mu + 1)(1 - q)^2 \mathcal{F}$  and  $B = (1 - q) \left[ \frac{1 - \mathcal{F}}{2} (1 + \mu) - \mu \right]$  are the same for both  $U_1$  and  $U_2$ . Clearly if  $A\sigma_{-i} + B > 0$ , then  $\sigma_i = 1$  will be optimal for player  $i$ , whereas  $\sigma_i = 0$  will be optimal for player  $i$  if  $A\sigma_{-i} + B < 0$ , and player  $i$  will be indifferent among all values of  $\sigma_i$  if  $A\sigma_{-i} + B = 0$ .

First, note that  $A = (\mu + 1)(1 - q)^2 \mathcal{F}$  must be positive, since  $\mu \in (0, 1)$ ,  $q \in (0, 1)$ , and  $\mathcal{F} \in (0, 1)$ . As such, keeping in mind that  $\sigma \in [0, 1]$ , it will be that  $A\sigma_{-i} + B > 0$  regardless of  $\sigma_{-i}$  – that is,  $\sigma_i = 1$  will be a strictly dominant strategy – when  $B > 0$ . Simple algebra employing the definition of  $B$  yields

$$B > 0 \iff \mathcal{F} < \frac{1 - \mu}{1 + \mu} \quad (6)$$

Because equation [5] holds for both players,  $\sigma_i = 1$  will be a dominant strategy for both players when the condition on  $\mathcal{F}$  given in equation [6] is satisfied. As such, the only Nash equilibrium profile in this regime is  $(\sigma_1, \sigma_2) = (1, 1)$ .

Further, since  $A$  is positive, it must also be that  $A\sigma_{-i} + B < 0$  regardless of  $\sigma_{-i}$  – that is,  $\sigma_i = 0$  will be a strictly dominant strategy – when  $B < -A$ . Simple algebra employing the definitions of  $A$  and  $B$  yields

$$B < -A \iff \left( q - \frac{1}{2} \right) \mathcal{F} > \frac{1 - \mu}{2(1 + \mu)} \quad (7)$$

When  $q \leq \frac{1}{2}$ , this inequality can never be satisfied, since the left hand side would not be positive but the right hand side must be (because  $\mu \in (0, 1)$ ). But for  $q > \frac{1}{2}$ ,

$$B < -A \iff \mathcal{F} > \frac{1 - \mu}{(1 + \mu)(2q - 1)} \quad (8)$$

Because equation [5] holds for both players,  $\sigma_i = 0$  will be a dominant strategy for both players when the condition on  $\mathcal{F}$  given in equation [8] is satisfied. As such, the only Nash equilibrium profile for these parameter values is  $(\sigma_1, \sigma_2) = (0, 0)$ . Note that this regime may be empty because  $\frac{1 - \mu}{(1 + \mu)(2q - 1)}$  may exceed 1. In particular, this fraction is less than 1 only when  $q > \frac{1}{1 + \mu}$ . Note that  $\mu \in (0, 1)$  implies that  $\frac{1}{1 + \mu} \in (\frac{1}{2}, 1)$ .

Now consider the range  $-A \leq B \leq 0$  that has not been discussed so far. When  $B$  falls within this interval,  $A\sigma_{-i} + B$  may be positive, negative, or zero, depending on the value of  $\sigma_{-i}$ . In particular, for specific values  $A^*$  and  $B^*$  of  $A$  and  $B$  respectively, there will be some cutoff value  $\sigma_{-i}^*$ . Because equation [5] holds for both players, the cutoff value  $\sigma_{-i}^*$  is the same for both players. In particular the definitions of  $A$  and  $B$  can be used to demonstrate that

$$A\sigma_{-i}^* + B = 0 \iff \sigma_{-i}^* = \left[1 - \frac{1 - \mu}{(1 + \mu)\mathcal{F}}\right] \frac{1}{2(1 - q)} \quad (9)$$

The expression for  $\sigma_{-i}^*$  is clearly an increasing function of  $\mathcal{F}$ . In particular, since  $B$  is decreasing in  $\mathcal{F}$ , when  $B = 0$ ,  $\mathcal{F}$  takes on its minimum value on  $-A \leq B \leq 0$ . For  $B = 0$ ,  $\mathcal{F} = \frac{1 - \mu}{1 + \mu}$ , and therefore  $\sigma_{-i}^* = 0$ . If  $q > \frac{1}{2}$ , when  $B$  is at its lower limit (when  $B = -A$ ),  $\mathcal{F} = \frac{1 - \mu}{(1 + \mu)(2q - 1)}$ , and therefore  $\sigma_{-i}^* = 1$ . If  $q \leq \frac{1}{2}$ , then from equation [7] it must be that  $B > -A$ , so this limit is never reached.

For  $\sigma_{-i} > \sigma_{-i}^*$ , the best response of player  $i$  will be  $\sigma_i = 1$ ; for  $\sigma_{-i} < \sigma_{-i}^*$ , the best response of player  $i$  will be  $\sigma_i = 0$ ; and for  $\sigma_{-i} = \sigma_{-i}^*$ , player  $i$  will be indifferent among all values of  $\sigma_i$ . As a consequence,  $(\sigma_1, \sigma_2) = (1, 1)$  will be a strict Nash equilibrium whenever  $\sigma_{-i}^* \in [0, 1)$ . That is, it is a strict Nash equilibrium for  $\mathcal{F} \in [\frac{1 - \mu}{1 + \mu}, \min(1, \frac{1 - \mu}{(1 + \mu)(2q - 1)})]$  when  $q > \frac{1}{2}$ ; and for  $\mathcal{F} \in [\frac{1 - \mu}{1 + \mu}, 1)$  when  $q \leq \frac{1}{2}$ . Also as a consequence,  $(\sigma_1, \sigma_2) = (0, 0)$  will be a strict Nash equilibrium whenever  $\sigma_{-i}^* \in (0, 1]$ . That is, it is a strict Nash equilibrium for  $\mathcal{F} \in (\frac{1 - \mu}{1 + \mu}, \frac{1 - \mu}{(1 + \mu)(2q - 1)})$  when  $q > \frac{1}{2}$ ; and for  $\mathcal{F} \in (\frac{1 - \mu}{1 + \mu}, 1)$  when  $q \leq \frac{1}{2}$ . In addition, there will also be a non-strict Nash equilibrium at  $(\sigma_1, \sigma_2) = (\sigma_{-i}^*, \sigma_{-i}^*)$  because the play of  $\sigma_{-i}^*$  by one's opponent makes one indifferent among all values of  $\sigma_i$ . That these are the only Nash equilibria is clear from the structure of the best

response functions.

As all of the Nash equilibria have now been characterized, it remains to determine whether or not each of these Nash equilibrium states corresponds to an evolutionarily stable state of the evolutionary game. Since each of the Nash equilibria is symmetric, the question reduces to whether or not the strategies specified in each equilibrium are evolutionarily stable.

Denote by  $u(x, y)$  the payoff to strategy  $x$  when confronted with strategy  $y$ . Then  $x$  is evolutionarily stable (Weibull 1995) if and only if  $u(x, x) \geq u(y, x) \forall y \neq x$  and  $u(x, x) = u(y, x) \implies u(x, y) > u(y, y)$ .

Notice first of all that if  $x$  is a strict Nash equilibrium against itself, then it must be evolutionarily stable, because  $u(x, x) > u(y, x) \forall y \neq x$ . As such, only the nonstrict Nash equilibria  $(\sigma_1, \sigma_2) = (\sigma_{-i}^*, \sigma_{-i}^*)$  must be checked for evolutionary stability. For these equilibria,  $u(\sigma_{-i}^*, \sigma_{-i}^*) = u(y, \sigma_{-i}^*) \forall y \neq \sigma_{-i}^*$ , so for  $\sigma_{-i}^*$  to be evolutionarily stable it must be that  $u(\sigma_{-i}^*, y) > u(y, y) \forall y \neq \sigma_{-i}^*$ . First suppose that  $\sigma_{-i}^* < 1$ . Then consider  $y = 1$ . When  $\sigma_{-i}^* < 1$ , it must be that  $y = 1$  is a strict Nash equilibrium. But this implies that  $u(y = 1, y = 1) > u(z, y = 1) \forall z \neq y$ . In particular this must hold for  $z = \sigma_{-i}^*$ . But this contradicts the requirement for  $\sigma_{-i}^*$  to be evolutionarily stable since there exists a  $y \neq \sigma_{-i}^*$  disobeying  $u(\sigma_{-i}^*, y) > u(y, y)$ . As such, none of the  $\sigma_{-i}^* < 1$  can be evolutionarily stable. Now take  $\sigma_{-i}^* = 1$ . A choice of  $y = 0$  and identical logic leads to the same conclusion. Thus, none of the nonstrict Nash equilibria are evolutionarily stable.

Combining the results from the different regimes  $B > 0$ ,  $B < -A$ , and  $-A < B < 0$ , we see that  $\sigma = 1$  is evolutionarily stable when  $\mathcal{F} \in (0, \min(1, \frac{1-\mu}{(1+\mu)(2q-1)}))$  for  $q > \frac{1}{2}$ , and when  $\mathcal{F} \in (0, 1)$  for  $q \leq \frac{1}{2}$ . Finally, we see that  $\sigma = 0$  is evolutionarily stable when  $\mathcal{F} \in (\frac{1-\mu}{1+\mu}, 1)$ . The set of evolutionarily stable states is therefore as given in the Proposition. ■

**Proof of Proposition 2.** We proceed using backwards induction. If given the opportunity, a good type of Adam would play  $C$  while a bad type of Adam would play  $D$  (both of these by definition). As a result, Eve's expected utilities for playing  $C$  and  $D$  will depend upon her posterior belief  $\bar{p}$  about Adam's type. Upon having received concessions  $x$ , her expected utility for playing  $C$  is  $\bar{p} + (1-\bar{p})(1-x_0+x-\epsilon)\beta_{Eve}$  while her expected utility for playing  $D$  is  $(1-x_0+x)\beta_{Eve}$ .

As such she strictly prefers to play  $C$  so long as  $\bar{p} > \frac{\epsilon\beta_{Eve}}{1-\beta_{Eve}(1-x_0+x-\epsilon)}$ , where the right hand side will be referred to as  $p^*(x)$ . Note that  $p^*(\tilde{x}) > p^*(0)$  because the right hand side is increasing in  $x$ .

We now consider the possible configurations of concession behavior by Adam, beginning with separating strategy profiles, which fully reveal Adam's type. For a profile in which the good Adam type chooses  $x = 0$  while the bad Adam type chooses  $x = \tilde{x}$ , the bad Adam type ultimately receives payoff  $(x_0 - \tilde{x})\beta_{Adam}^b - c$  because Eve knows him to be bad and plays  $D$ , while by choosing  $x = 0$  he could have earned a payoff  $(x_0 + \epsilon)\beta_{Adam}^b$  (saving on the concession and fooling Eve about his type in the process). So there is clearly no separating equilibrium of this kind. For a profile in which the good Adam type chooses  $x = \tilde{x}$  while the bad Adam type chooses  $x = 0$ , the good Adam type gets payoff  $1 - c$  because Eve knows him to be good and plays  $C$ , while by choosing  $x = 0$  he would have earned payoff  $x_0\beta_{Adam}^g$ . Meanwhile the bad Adam type gets payoff  $x_0\beta_{Adam}^b$ , while by choosing  $x = \tilde{x}$  he would have earned payoff  $(x_0 - \tilde{x} + \epsilon)\beta_{Adam}^b - c$ . As such, neither type has an incentive to deviate, and there is therefore a PBE of this kind, so long as  $(\epsilon - \tilde{x})\beta_{Adam}^b \leq c \leq 1 - x_0\beta_{Adam}^g$ .

Now consider pooling strategy profiles, which do not fully reveal Adam's type. Throughout we assume that off-equilibrium-path beliefs are intuitive, in the sense that deviants to out-of-equilibrium strategies are assumed to be of the type more naturally amenable to the deviation (ie, that a deviation to  $x = 0$  from pooling on  $x = \tilde{x}$  leads to  $\bar{p} = 0$  – a belief that the deviant is of the bad type – whereas deviation to  $x = \tilde{x}$  from pooling on  $x = 0$  leads to  $\bar{p} = 1$  – a belief that the deviant is of the good type). For a profile in which both Adam types choose  $x = \tilde{x}$ , there are two cases. If  $p_0 > p^*(\tilde{x})$ , Eve plays  $C$  on the equilibrium path. A good Adam type therefore gets payoff  $1 - c$  on the equilibrium path but  $x_0\beta_{Adam}^g$  by deviating, whereas a bad Adam type gets payoff  $(x_0 - \tilde{x} + \epsilon)\beta_{Adam}^b - c$  on the equilibrium path but  $x_0\beta_{Adam}^b$  by deviating. If on the other hand  $p_0 < p^*(\tilde{x})$ , Eve plays  $D$  on the equilibrium path; a bad Adam type therefore gets payoff  $(x_0 - \tilde{x})\beta_{Adam}^b - c$  on the equilibrium path but a higher value  $(x_0)\beta_{Adam}^b$  by deviating. So an equilibrium pooling on  $x = \tilde{x}$  is possible if and only if  $p_0 \geq p^*(\tilde{x})$  and  $c \leq \min[1 - x_0\beta_{Adam}^g, (\epsilon - \tilde{x})\beta_{Adam}^b]$ . For a profile in which both Adam types choose  $x = 0$ , there are also

two cases. If  $p_0 > p^*(0)$ , Eve plays  $C$  on the equilibrium path. A good [bad] Adam type therefore gets payoff 1 [ $(x_0 + \epsilon)\beta_{Adam}^b$ ] on the equilibrium path but a clearly inferior payoff by deviating (which would amount to paying a cost in order to be disbelieved by Eve), so clearly neither actor has an incentive to deviate by the definitions of the types. Finally if  $p_0 < p^*(0)$ , Eve plays  $D$  on the equilibrium path. A good Adam type therefore gets payoff  $x_0\beta_{Adam}^g$  on the equilibrium path but  $1 - c$  by deviating, whereas a bad Adam type gets payoff  $x_0\beta_{Adam}^b$  on the equilibrium path but  $(x_0 - \tilde{x} + \epsilon)\beta_{Adam}^b - c$  by deviating. So an equilibrium pooling on  $x = 0$  is possible if and only if either (i)  $p_0 \geq p^*(0)$  or (ii)  $p_0 < p^*(0)$  and  $c \geq \max[1 - x_0\beta_{Adam}^g, (\epsilon - \tilde{x})\beta_{Adam}^b]$ . ■

**Proof of Proposition 3.** Eve's posterior beliefs are relevant only insofar as they affect her decision to play  $C$  or  $D$  in the trust game. As such, any posterior belief above her threshold level of trust will lead her to play  $C$  and any belief below that threshold will lead her to play  $D$ . As such, the possible postures can be partitioned into four categories: those which will lead Eve to play  $C$  no matter what concessions Adam offers; those which lead Eve to play  $D$  no matter what; those which lead Eve to play  $D$  iff Adam offers  $x = 0$ ; and those which lead Eve to play  $C$  iff Adam offers  $x = 0$ . These posture categories will be represented by their limiting members, respectively:  $(\bar{p}(0), \bar{p}(\tilde{x})) = (1, 1), (0, 0), (0, 1), (1, 0)$ .

We now consider an evolutionary game in which different posture types evolve for fixed values of the parameters of the model. Because the Eve role is the only one involving the types being selected for, it is clear that any posture achieving optimal payoffs will be an equilibrium outcome of the evolutionary game.

Suppose that  $p^*(\tilde{x}) \geq p_0 \geq p^*(0)$  and that  $c \leq \min[1 - x_0\beta_{Adam}^g, (\epsilon - \tilde{x})\beta_{Adam}^b]$ . Then by Proposition 2 there is one type of equilibrium in the basic reassurance game: a pooling equilibrium on  $x = 0$ . As such there is one category corresponding to Bayesian postures (given the assumption about intuitive off-equilibrium-path beliefs):  $(\bar{p}(0), \bar{p}(\tilde{x})) = (1, 1)$ . The posture  $(1, 1)$  (and others in its category) yields for Eve an expected payoff of  $p_0[1] + (1 - p_0)[(1 - x_0 - \epsilon)\beta_{Eve}]$  because neither Adam type makes a concession and the bad Adam type takes advantage of the cooperative gesture she automatically makes. Can a non-Bayesian posture induce a better outcome? A posture of the category  $(\bar{p}(0), \bar{p}(\tilde{x})) = (1, 0)$  yields the same payoff because it results in the same action profiles;

both Adam types will choose  $x = 0$ , Eve will trust them anyway, and the bad Adam type will betray that trust. A posture of the category  $(0, 0)$  will yield an expected payoff of  $(1 - x_0)\beta_{Eve}$ ; both Adam types will choose  $x = 0$ , and Eve will not trust them. Finally, consider a posture of the category  $(0, 1)$ . Given that  $c \leq \min[1 - x_0\beta_{Adam}^g, (\epsilon - \tilde{x})\beta_{Adam}^b]$ , the good type will prefer  $x = \tilde{x}$  over  $x = 0$  (because  $1 - c > x_0\beta_{Adam}^g$  follows) while the bad type will prefer  $x = \tilde{x}$  over  $x = 0$  as well (because  $(x_0 - \tilde{x} + \epsilon)\beta_{Adam}^b - c > x_0\beta_{Adam}^b$  follows). As such Eve's expected payoff will be  $p_0[1 + c] + (1 - p_0)[(1 - x_0 - \epsilon + \tilde{x})\beta_{Eve} + c]$ . But this clearly dominates the expected payoffs from the other postures. So the set of equilibrium postures consists of all postures of category  $(0, 1)$ , all of which are biased postures.

Now suppose that  $p_0 \geq p^*(0)$  and that  $(\epsilon - \tilde{x})\beta_{Adam}^b \leq c \leq 1 - x_0\beta_{Adam}^g$ . Then by Proposition 2 there are two types of equilibrium in the basic reassurance game: the separating equilibrium and a pooling equilibrium on  $x = 0$ . As such there are two categories corresponding to Bayesian postures (given the assumption about intuitive off-equilibrium-path beliefs):  $(1, 1)$  and  $(0, 1)$ . Postures in the categories  $(1, 1)$ ,  $(1, 0)$ , and  $(0, 0)$  yield the same payoffs as they did in the previous paragraph. But a posture of category  $(0, 1)$  now induces the good Adam type to choose  $x = \tilde{x}$  but the bad Adam type to choose  $x = 0$  because  $(\epsilon - \tilde{x})\beta_{Adam}^b \leq c \leq 1 - x_0\beta_{Adam}^g$ . As a result, Eve's expected payoff will be  $p_0[1 + c] + (1 - p_0)[(1 - x_0)\beta_{Eve}]$ , which clearly dominates the expected payoffs from the other postures. So the set of equilibrium postures consists of all postures of category  $(0, 1)$  which are a proper subset of the Bayesian postures. ■

## Notes

<sup>1</sup>Rational choice models, of course, can serve many different masters: normative theory, the development of behavioral prescriptions, the establishment of benchmark outcomes against which experimental results can be compared (Schotter 2005), and others. The exclusive focus of this paper will be the use of rational choice models in positive political theory, that is, to explain some set of outcomes or phenomena in the political world.

<sup>2</sup>A useful summary of some of the vast literature on such violations can be found in Camerer (1995).

<sup>3</sup>It is important to note that, throughout the paper, the word “beliefs” will be construed specifically in the sense of beliefs about some uncertain state of the world, rather than in any way that might be confused or conflated with preference. That is, the word “beliefs” will be employed to describe individuals’ subjective views on what are essentially *factual* questions.

<sup>4</sup>It should be stressed at an early stage that this paper does not advocate a view that individuals choose in any conscious sense how to update beliefs, or what kind of cognition they would like to possess more generally. A discussion of the interpretation favored here will follow at the end of this section and in the following section.

<sup>5</sup>It is important to distinguish between two conceptually distinct approaches to modelling relaxations of the standard rationality assumptions. In the first type, “behavioral” features, generally inspired by stylized facts from psychology or from economics experiments, are exogenously assigned to actors. In general, the purpose of such studies is not to investigate the possible origins of such deviations from rational choice behavior, but rather to infer the potential consequences of such deviations for economic or political phenomena of interest. For example, Rabin and Schrag (1999) develop a model in which they assume players exhibit a particular kind of confirmatory bias, and study the long-term implications of this bias for information aggregation; Hafer and Landa (2005) develop a model in which agents are assumed to exhibit a violation of negative introspection, and examine the implications for the dynamics of deliberation; and Dickson and Scheve (2006) assume that voters are susceptible to ethnic priming by politicians’ rhetoric, and then examine the implications of such susceptibility for various aspects of political competition, such as the policy platforms that can be chosen by candidates in electoral equilibrium. In contrast, the approach advocated here is to allow “behavioral” features to emerge endogenously from the strategic structure of the situation in which actors find themselves.

<sup>6</sup>In the same way that models taking group behavior or characteristics as primitives are frequently susceptible to charges of arbitrariness or indiscipline, individual deviations from rational behavior have traditionally been viewed with similar kinds of suspicion. Because the equilibrium framework is maintained here, and because belief updating is considered endogenously, a high degree of discipline is maintained – actors can form beliefs only in ways that are consistent with an equilibrium profile. That the model’s predictions involve *equilibrium* outcomes automatically guarantees that the ways in which actors form beliefs in accordance with those predictions is automatically credible.

<sup>7</sup>The literature most closely related to the approach taken here is the growing literature on endogenous preferences (Acemoglu and Yildiz (2001); Heifetz and Segev (2004); Heifetz, Shannon and Spiegel (2004a and 2004b); Ok, Kockesen, and Sethi (2000)). In this literature, players are allowed to perceive (and act on) preferences that can deviate from their underlying actual preferences – that is, the payoffs which determine the fitnesses of different types in an underlying evolutionary game. The basic findings of this literature are that, in many settings, agents will evolve perceived preferences that do indeed differ from the underlying fitness preferences. In terms of evolution of beliefs as distinct from preferences, the only paper of which we are aware is Gehrig, Guth, and Levinsky (2004), which considers the evolution of beliefs in a market.

<sup>8</sup>While the set of possible outcome states is unaffected by the details of the underlying dynamics so long as they are payoff-monotone, other quantities of potential interest, such as the time required for the system to approach equilibrium, and the probability that a given outcome will be selected, will depend on these details.

<sup>9</sup>While the language employed here, and by evolutionary game theorists more generally, often evokes images of genetic hardwiring, the reader is reminded that there always exists a parallel story of social adaptation by boundedly-rational actors. Thus, a boundedly-rational actor might be said to be “programmed” to behave in a certain way if she just gets it in her head to behave that way.

<sup>10</sup>In biological applications for which genetic hardwiring is operative, “genes” for strategies will come and go due to the differential reproductive and mortality rates of their carriers. In applications involving social adaptation, “memes,” the social counterpart of genes that can loosely be thought of as ideas, will come and go over time because people abandon practices that perform poorly and adopt new ones that do better, through imitation, reinforcement learning, or some other boundedly-rational adaptive process.

<sup>11</sup>In the biological context, “mutation” generally means mutation of a gene. In the social context, “mutation” corresponds to an individual innovation, or to a mistake or other random act.

<sup>12</sup>It should be emphasized that the model in this section makes no attempt to “rationalize” perceptual biases or other forms of non-Bayesian behavior in the context of individual decision problems that are unrelated to group decision making. However, in fields like public opinion research, in which the preferences and beliefs of mass publics are the objects of study, few if any of the challenges faced by actors are genuinely individual in nature. Just as recent formal literatures on juries and voting have stressed the importance of “strategic effects” on outcomes and the out-of-equilibrium nature of some “sincere” behavior, this paper argues that it is improper to use decision-theoretic constructs like Bayes’ Rule, which can be thought of as “sincere” belief-formation behavior, in the deeply strategic context of group decision making. Further, in other areas of application, many choices that are often thought of as being individual in nature may, in the larger scheme of things, actually have an implicit group-based component. Coordination problems are such a constant presence in human interactions that relatively few of the opinions held by individuals seem likely ultimately to remain relevant to purely individual choice alone.

<sup>13</sup>It may seem strange that the model includes a noiseless signal, since perceptual biases are

often thought of as arising in settings of incomplete information. If it can be shown, however, that perceptual biases can be expected in equilibrium *even when signals are not noisy*, this strengthens the case that Bayesian models should be rejected not only on empirical, but also on theoretical grounds.

<sup>14</sup>While the model described here is written in terms of a perceptual bias, it could just as well have been set in alternative terms as a model of non-Bayesian belief updating with unbiased perception. For example, rather than saying that an actor confronted with a signal contradicting her prior beliefs incorrectly perceives the world with probability  $1 - \sigma$ , but then updates her beliefs based on that perception using Bayes' Rule, one might alternatively speak of an individual who perceives the world accurately, but who in updating her beliefs draws the opposite inference from the prescription of Bayes' Rule with probability  $1 - \sigma$ . For the purposes of this section, these alternative frames are essentially interchangeable, and phrases like "perceptual biases" and "non-Bayesian updating" are used as synonyms. There may, however, exist other settings in which it is important that distinctions be drawn with greater care.

<sup>15</sup>Of course, the relative ease with which coordination can be achieved in any concrete instance will be influenced, perhaps among other factors, by the degree to which players are free to communicate; the players' beliefs not only about which option is best, but about the other player's beliefs, the other player's beliefs about one's own beliefs, and so on; and any relevant history of interaction between the individuals. What will really be essential here is not that the probability of coordination (on the Pareto-dominant focal point) is 1, but rather that the probability of coordination is high enough relative to that which obtains when perceived preferences diverge and no such clear focal point exists. Still, a coordination game with a Pareto-dominant, payoff-dominant, risk-dominant outcome does possess an exceptionally clear focal point. Particularly in a setting of full communication, it would seem perverse for players to settle on anything apart from this outcome.

<sup>16</sup>The treatment of  $\mathcal{F}$  as an exogenous variable may seem strange, since the actual choices made by individuals are generally the primary objects of interest in game theoretic models. One conceptualization of  $\mathcal{F}$  that may reassure some readers is to think of its value as being determined by the probabilities with which players choose different strategies in the mixed-strategy equilibrium of the perceived game; because the situation is symmetric, there is no obvious reason to choose one of the pure strategy equilibria over the other, and it seems behaviorally plausible that the mixed strategy equilibrium will be played much more commonly in this setting than it would be in the perceived game in which players have aligned preferences. And of course, the mixed strategy equilibrium involves a positive probability of coordination failure. Another intuition that may serve to justify the exogeneity of  $\mathcal{F}$  is that individuals in real-life group interactions are typically confronted by a constant stream of novel coordination problems in going about group business. In evolutionary game theory terms, the novelty of these problems would make it impossible for selection to take place at the level of surface strategies themselves; as a result, it is reasonable to imagine selection taking place over perceptual strategies, and to treat play in novel coordination settings with multiple equilibria as empirical distributions.

<sup>17</sup>When players have divergent perceived preferences, they find themselves in a battle-of-the-sexes scenario, in which it seems reasonable to assume that coordination is more difficult than

it is in the setting when perceived preferences are aligned, because battle-of-the-sexes games of the kind posited here do not possess a focal point of comparable strength. An alternative way of thinking about this is that the battle of the sexes game is essentially a kind of bargaining problem; and in the literature on bargaining “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” and this is true “even when it is evident that there are gains to be had from agreement.” (Roth 1995) Biased perception can be thought of as a way of bypassing these risks inherent in the bargaining process.

<sup>18</sup>Dickson (2006a) contains results concerning the evolutionarily stable perceptual states that exist under other circumstances, as well as some generalizations of the model presented here, including the effects of intergroup competition on within-group perception formation. In each of the other circumstances and generalizations, it is found, as in the Proposition here, that there are conditions (i.e., values of  $q$ ,  $\mu$ , and  $\mathcal{F}$ ) under which there exist no equilibria involving entirely Bayesian populations, and in which confirmatory bias-like deviations from Bayesian updating do exist in equilibrium.

<sup>19</sup>Unless, of course, there exist constraints of a sort that are not modelled here. For example, it might be optimal for an individual to have different heights under different conditions, but a given (adult) individual is physically constrained to have a fixed height. The degree to which individual perceptual behavior might be responsive or constrained is unclear both theoretically and empirically, but it seems reasonable to expect that such behavior could be context-sensitive at least to some degree.

<sup>20</sup>The fact that the equilibria mostly involve boundary values may simply be a reflection of the discrete nature of the coordination problem modelled here. It seems intuitive that, if a continuum of coordination states existed rather than a discrete pair, then some intermediate level of bias might be optimal in many circumstances, as long as the probability of coordination failures is an increasing function of the “distance” between individuals’ perceived ideal coordination states and individuals’ utilities over outcomes is a decreasing function of the “distance” between the coordination state selected and perceived ideal coordination states.

<sup>21</sup>This depends, of course, on the details of the coordination environment as embodied in  $\mathcal{F}$ . For the assertion to be true, it must be the case, in terms of the  $\mathcal{F} - \mu$  tradeoff, that disagreement on the minor issue matters more to the prospects for coordination than the minor issue matters to individuals. It seems intuitive that this will frequently be the case. In the simple two-person models solved here, the replacement of unanimity with symmetric conflict would seem likely to pose a non-negligible difficulty — or at least a difficulty that is non-negligible compared with individuals’ potential issue losses on a issue that matters very little to them.

<sup>22</sup>Because the topics are basically unrelated, individuals’ true preferences over the issues should be uncorrelated. That their perceived preferences are not should be seen as a puzzle that calls for explanation.

<sup>23</sup>Of course, as indicated above, arguments about the “rationality of irrationality” are familiar in the international relations literature (see eg Zagare 1990). A distinguishing feature of the framework advocated here is that the irrationalities that actors may demonstrate are not arbitrary

or idiosyncratic in nature, but rather are constrained by being endogenously determined within the context of the equilibrium of a game – and are therefore, in a specific sense, not only less arbitrary than “irrational” beliefs that do not emerge endogenously, but also automatically credible because they are part of an equilibrium behavioral profile. The beliefs formed in this way will be referred to from time to time as “in-equilibrium non-Bayesian beliefs” in order to emphasize this feature of the model.

<sup>24</sup>For expositional clarity, the model is presented in terms of two unitary actors, to be referred to using the female and male personal pronouns respectively. However, the insights communicated by the model are intended to extend to belief formation in mass publics – for example, the beliefs of ordinary citizens about the intentions of a state in conflict with their own, or about the nature of individuals who are members of ethnic, national, or religious groups in conflict with their own. Indeed, in many – though certainly not all – settings it seems natural to imagine that elites are closer approximations to classical rational actors than are members of mass publics, but that mass opinion shapes the incentives of elites and the set of actions from which elites may feasibly choose.

<sup>25</sup>Where  $p^*(\tilde{x}) > p^*(0)$  are both threshold values defined explicitly in the Appendix.

<sup>26</sup>As such, “Eve” is more precisely thought of as a label for a role rather than as a specifically incarnated actor.

<sup>27</sup>Because in some situations there may be multiple Perfect Bayesian Equilibria indicating different possible posterior beliefs upon observation of the same action – depending on the specific nature of the common conjecture shared by actors in a given equilibrium – there may therefore be more than one possible “Bayesian posture” in a given situation. In a Perfect Bayesian Equilibrium, equilibrium beliefs are dictated by equilibrium strategies, whereas a posture as defined here is “strategy-like” in its nature. In a particular sense, adopting a posture is a more straightforward act than the process of belief formation in a standard Perfect Bayesian Equilibrium; anyone who has seen undergraduates struggle to understand how the same action can have different belief consequences in Perfect Bayesian Equilibria may find this straightforwardness behaviorally appealing.

<sup>28</sup>That is, compare equilibrium outcomes across models when all of the values of all of the parameters are held fixed.

<sup>29</sup>In addition, one might argue that this sort of mini-ultimatum – “if you don’t do  $x$  I won’t trust you any more...” – seems to resonate more strongly with the psychology of arguments in everyday life than does the more tortured logic of Perfect Bayesian Equilibrium.

<sup>30</sup>Indeed, the idiom of “subscribing” to an ideology in itself implicitly hints at a loss of individual capacity in seeing the world clearly for oneself; once you’re signed up, whatever you get in the mail is what you have to read.

<sup>31</sup>Although it is tempting to think of these different extremes as somehow mapping onto differing paradigms of biological versus social adaptive processes, the correspondence is not necessarily as neat as it might appear at first glance. For example, while one would be inclined to think of

biologically hard-wired cognitive properties as being in a specific sense more rigidly defined than those which might be the result of social learning, even genuinely hard-wired behaviors might be programmed in a conditional way: think like  $x$  if  $a$  but think like  $y$  if  $b$ . Of course, the interaction between genetics and the environment is so intricate that thinking about this distinction in this way might be a fool's errand in the first place.

## References

- Acemoglu, Daron and Muhamet Yildiz (2001) 'Evolution of Perceptions and Play', MIT Mimeo.
- Bawn, Kathleen (1999) 'Constructing "Us": Ideology, Coalition Politics, and False Consciousness', *American Journal of Political Science* 43/2: 303-34.
- Camerer, Colin (1995) 'Individual Decision Making' in J. Kagel and A. Roth (Eds.), *Handbook of Experimental Economics*. Princeton: Princeton University Press.
- Camerer, Colin (2003) *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton: Princeton University Press.
- Cox, Gary (1997) *Making Votes Count*. Cambridge UK: Cambridge University Press.
- Dickson, Eric S. (2006a) 'Perceptual Biases, Non-Bayesian Beliefs, and Coordination: A Model of Preference Formation', NYU Mimeo.
- Dickson, Eric S. (2006b) 'A Model of Endogenous Suspicion in Conflict', NYU Mimeo.
- Dickson, Eric S. and Kenneth Scheve (2006) 'Social Identity, Political Speech, and Electoral Competition', *Journal of Theoretical Politics* 18/1: 5-39.
- Fudenberg, Drew and David Levine (1998) *The Theory of Learning in Games*. Cambridge MA: MIT Press.
- Gerber, Alan and Donald P. Green (1998) 'Rational Learning and Partisan Attitudes', *American Journal of Political Science* 42/3: 794-818.
- Green, Donald, Bradley Palmquist, and Eric Schickler (2002) *Partisan Hearts and Minds*. New Haven: Yale University Press.
- Gehrig, Thomas, Werner Guth and Rene Levinsky (2004) 'The Commitment Effect in Belief Evolution', *Economics Letters* 85/2: 163-6.
- Hafer, Catherine and Dimitri Landa (2005) 'Deliberation as Self-Discovery and Institutions for Political Speech', NYU Mimeo.
- Hastorf, Albert and Hadley Cantril (1954) 'They Saw a Game: A Case Study', *Journal of Abnormal and Social Psychology*.

- Heifetz, Aviad and E. Segev (2004) 'The Evolutionary Role of Toughness in Bargaining', *Games and Economic Behavior* 49: 117-34.
- Heifetz, Aviad, Chris Shannon, and Yossi Spiegel (2004a) 'The Dynamic Evolution of Preferences', working paper.
- Heifetz, Aviad, Chris Shannon, and Yossi Spiegel (2004b) 'What to Maximize if You Must', working paper.
- Kydd, Andrew (2000) 'Trust, Reassurance and Cooperation', *International Organization* 54/2: 325-57.
- Ok, Efe, L. Kockesen and R. Sethi (2000) 'The Strategic Advantage of Negatively Interdependent Preferences', *Journal of Economic Theory* 92: 274-99.
- Page, Benjamin I. and Robert Y. Shapiro (1992) *The Rational Public*. Chicago: University of Chicago Press.
- Rabin, Matthew and Joel Schrag (1999) 'First Impressions Matter: A Model of Confirmatory Bias', *Quarterly Journal of Economics* 114/1: 37-82.
- Roth, Alvin E. (1995) 'Bargaining Experiments' in *The Handbook of Experimental Economics*, John H. Kagel and Alvin E. Roth, eds. Princeton: Princeton University Press.
- Schelling, Thomas (1960) *The Strategy of Conflict*. Cambridge MA: Harvard University Press.
- Schotter, Andrew (2005) 'Strong and Wrong: The Use of Rational Choice Theory in Experimental Economics', NYU Mimeo.
- Tajfel, Henri (1981) *Human Groups and Social Categories*. Cambridge UK: Cambridge University Press.
- Vega Redondo, Fernando (1996) *Evolution, Games, and Economic Behavior*. Oxford: Oxford University Press.
- Weibull, Jorgen (1995) *Evolutionary Game Theory*. Cambridge MA: MIT Press.
- Zagare, Frank C (1990) 'Rationality and Deterrence', *World Politics* 42/2: 238-60.