Summer 9-1-2014

Making Sense of Unexpected Preferences

Gordon Alexander Arsenoff
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Political Science

Dissertation Examination Committee:
Randall Calvert, Chair
Jeff Gill
Gary Miller
John Patty
Maggie Penn
Brian Rogers

Making Sense of Unexpected Preferences

by

Gordon Arsenoff

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2014

St. Louis, Missouri

Contents

Acknowledgments

Thanks are in order first and foremost to my advisor Randall Calvert and to my committee members Jeff Gill, Gary Miller, John Patty, Maggie Penn, and Brian Rogers.

Data used in the ideal point estimation chapter were collected and generously shared by Adam Bonica, Andrew Martin, Harold Spaeth, and Thomas Hansford. Special thanks are in order to Prof. Hansford for supplying me with the data used on amicus curiae briefs by organized interests, which were not in publication at the time of my dissertation defense.

I am grateful for the support of my mother Elizabeth Kuhne, as well as of my stepfather Robert Bindschadler and my sister Sarah Arsenoff.

My best friend in the dissertation-writing period has been Prof. SSgt Kate Eakin. My best co-worker in dissertation writing has been Jessica Ruthven.

Helpful commenters on these essays besides those persons mentioned above have included Rob Alba, Rick Blanke, Frederic Bush, Chris Chiego, Ron Freiwald, Morgan Hazelton, Chad Henson, Tasos Kalandrakis, Walter Mebane, Michael Nelson, Irina Panovska, Jesse Richman, Keith Schnakenberg, John W. Simon, Anthony Stenger, Jessica Stoll, and Jerry Vinokurov.

Finally, I have benefited greatly from the assistance of the staff of the Department of Political Science, including Heather Sloan-Randick during my whole period of graduate study, Colleen Daues and Sue Tuhro as of the defense of this dissertation, and others.

ABSTRACT OF THE DISSERTATION

Making Sense of Unexpected Preferences

by

Gordon Arsenoff

Doctor of Philosophy in Political Science

Washington University in St. Louis, 2014

Professor Randall Calvert, Chair

This dissertation includes three papers using quantitative models to sensibly describe what kinds of preferences political actors will or actually do hold when existing theory offers no insight. The first two papers use evolutionary game theory to predict ways in which politicians, artificially selected on the basis of good performance to remain in office, will in the long run diverge from instrumental rationality as ordinarily assumed in game theory. The first sets out a general principle for producing models of preference evolution in games as political models, namely, that the information about opponent preferences necessary for evolution of non-rational preferences comes from opponents' previous plays, and applies it to two simple games. The second uses the same principles in more detail on a bargaining game that models the plea negotiations between a prosecutor and a defense attorney, leading to a conclusion that failure to learn from setbacks during a trial is an evolutionarily favored trait among prosecutors. The third paper addresses the ideological preferences of Supreme Court justices, which existing statistical models do not effectively compare to those of elected officials since the two groups never vote on the same items, by identifying a set of political actors with whom both groups commonly interact: organized interest groups who vote on Supreme Court cases with *amicus curiae* briefs and on electoral candidates using campaign donations.

Chapter 1

Introduction

The successes of positive political theory have largely been achieved by modeling political actors, whether at the individual or group level as dictated by the scale of the political phenomena being modeled, as goal-driven agents seeking their own well-being through interaction with others in the political sphere. The term "rationality" is used to encompass two connected characters of the players in formal models. One is the holding of well-defined preferences over outcomes, especially, where meaningful, ones that match up with more objective rewards, such as profit or the achievement or retention of office. The other is the capacity to compute and follow through on the strategies for interaction that maximize the payoffs they can expect to get with respect to their preferences. Predictions made through approximating political participants as unboundedly rational game players, as defined above, serve to explain very much of the behavior actually observed among those participants.

Empirical studies of human beings in laboratory settings consistently confirm, meanwhile, that such rationality assumptions are not literally true among human subjects. In principle, at least some further variation in political choices could be explained if models also included variation in players' rationality characteristics. However, modeling players as compromised or bounded in their rationality may severely compromise models' math-

ematical tractability, such that the extra effort needed to do so may be seen as not worth the gains that would result in predictive accuracy. More importantly (especially at a time where computing power is sufficient to obtain numerical or simulated solutions to many intractable formal theory problems), there is a question of what alternative sets of preferences can be assumed with mathematical rigor rather than ad hoc, and in what sense theory does any explanation if actions are to be explained by assuming that participants hold preferences that are deduced from empirical observation of their actions. Indeed, it is intuitive that, even though real persons may fall more or less short of instrumental rationality, those who actually participate in politics rather than seeking or being forced to exit will be those whose motivations are closest to single-minded office-seeking and whose cognitive capacity to act on those motivations is least constrained. The validity of such an intuition, of course, is a natural subject for evolutionary game theory.

In this dissertation, I propose and implement ways for positive political theorists to account also for variation in preferences themselves, including those that define rationality, among those who participate, or may be selected to participate, in political processes, both in general and in some specific cases, without abandoning the tractability or rigor necessary to make predictions in the first place.

In chapter 2, I import into political science a general framework from the economics literature for studying the evolution of preferences themselves, usually referred to as the *indirect evolutionary approach*, and I supply that general framework with a politically plausible and mathematically representable microfoundation for an assumption on which its results of interest critically depend. According to results derived using the indirect evolutionary approach, many kinds of non-rational, non-fitness-maximizing preferences can emerge and thrive in evolutionary competition, as long as players get information about what kinds of preferences their opponents do hold. Through what channels opponents get information about each other's behavioral types, though, is left unstated. In those cases where such a channel is suggested, it is typically something like an assumption that

players signal their types involuntarily in a way that is costly to suppress; such an assumption, however, itself asks for evolutionary justification. I add the principle that players can in fact acquire such information about their opponents by considering what those opponents have done in past interactions of the same kind. While politicians' mental states likely cannot be directly observed by those interacting with them, what they have done during their careers is much more commonly known to the public. Using the indirect evolutionary approach together with the principle of preference disclosure through history of play, I derive results about the long-run rationality characteristics of players in several cases: first in any of a large class of games, then in a specific game within the class, and last in another game outside the class, where precision is limited due to tractability but substantive insight is still provided.

Political players may exhibit non-material preferences implicitly by forming new beliefs in ways inconsistent with probability theory when new information is revealed. Put another way, players may simply hold preferences over what to believe. In chapter 3, I use the same principles as in the previous chapter to analyze in more depth a particular kind of interaction of political concern – the process of plea bargain negotiation between prosecutors and defense attorneys – and how it can be expected to shape the way repeat participants in such negotiations learn from events in the courtroom between offers. Specifically, I model trials as alternating rounds of bargaining and fighting between a prosecutor and a defense attorney over the sentence ultimately assigned to the defendant, where the defense attorney knows more than the prosecutor does about how hard it will be to get a conviction. Using this model, I show that a prosecutor who does not update his belief about the strength of his case following setbacks in courtroom fights can expect to end up winning longer sentences overall than one who updates in a rational, Bayesian fashion, as long as the defense knows about the prosecutor's learning style. Since the deals the prosecutor has struck in previous cases are matters of public record and will differ between biased and rational prosecutors, the defense will at least sometimes have such

knowledge about the prosecutor's rationality. Since, furthermore, winning longer sentences is connected with career furtherance for prosecutors, we should expect to see the kind of optimism bias described in those who actually fill the role of prosecutor. (None of the games treated in the previous chapter are bargaining games as this one is.)

One kind of preference frequently distinguished from instrumental rationality is ideology. Although elite actors' ideologies may sometimes be rational adoptions of their constituents', ideology can also be understood as a preference for policies other than those maximizing material rewards. The latter characterization applies especially among citizens at large and the interest groups into which they organize themselves, and among politicians, like Supreme Court justices, whose continuance in office is not dependent on voters and voters' ideologies. In chapter 4, I demonstrate a method for measurement of ideology in a consistent way across political bodies whose members do not interact with each other or take positions on the same questions. I exploit the fact that organized interest groups take ideologically-driven positions both on Supreme Court cases through amicus curiae briefs and on candidates for Congress and the Presidency through campaign donations. Existing models relate Justices' ideologies to their votes on cases and relate the ideologies of candidates to their donors, but the treatment of amici as additional Supreme Court voters that allows unification of these models is novel. Ideology scores on a common scale for legislative-, executive-, and judicial-branch actors will improve researchers' ability to predict the outcomes of interactions between two or more branches.

Chapter 2

I Saw Who You Are And I Know What You Did

2.1 Introduction

In formal theory, interpretations of institutions' importance lie generally on a spectrum with two poles: one understanding institutions as "the rules of the game" and another treating them as equilibria emerging from play of a game outside institutions themselves (Calvert, 1995). In the latter, institutions are effects of political realities, but in the former they are considered of interest for how they function, together with other political realities, as causes of political outcomes. An institution is modeled as a game with known rules, and the political output of the institution is predicted to resemble the "final score" of the game, when played by contestants with certain characteristics like a set of traits described as "rationality".

This approach may give institutions too little credit, however. Can institutions themselves cause the characteristics of participants, including limits on their ability to play rationally? And, if so, can formal theory still be of use in studying them, including studying what kinds of deviation from rationality they will favor? This chapter will argue that

institutions where repeat participants are more likely to remain participants if they realize higher material rewards, and in which participants know about the behavior on previous dates of those with whom they interact, will come to be characterized by participants who play in ways other than ones that maximize their own objective payoffs.

How can deviations from rationality possibly be favored by artificial selection in a political institution where better payoffs make players more likely to play again? The key is that players will learn what to expect from opponents using those opponents' histories of play, if any. Players' reputations can induce changes in play from their opponents (relative to what their opponents would do against payoff-maximizing players) that may increase expected payoffs more than they are directly decreased by making non-payoff maximizing moves. A behavioral type of player for whom such net reputational benefits hold when playing against the same behavioral type will thrive compared to strict payoff-maximizers.

This account contrasts with ones like that of Frank (1988), who argues that the emotions driving non-rational preferences will generally be associated with observable physical cues that give opponents knowledge of each other's types. This chapter thus helps expand the set of contexts in which we can expect to see deviations from rationality survive to include political contexts where strategic interactions occur between players not in close enough physical proximity to observe each other's emotional "tells", to non-payoff-maximizing preferences not connected with strong moral sentiments, and to the plausible scenario in which those who participate in politics are disproportionately psychopaths who do not exhibit typical side effects of emotion.

This chapter also complements previous accounts relying on reputation as a condition under which payoff-maximizing behavior will not go away. Where Kreps and Wilson (1982) suggest that allowing players to have a reputation for irrational (non-payoff maximizing) play allows rational types to get ahead by mimicking irrational ones, this chapter argues that those who get ahead through reputations for irrational play need not be rational themselves; even if not characterized by the ability to defect and revert to payoff-

maximizing play, those who play so as to maximize functions other than their material payoffs will, in general, be those left playing the game in the long run. Kreps and Wilson (1982) require the existence of at least some players of an intrinsically tough type to support continued tough behavior by players of normal type, but they lack a justification for the existence of intrinsically tough players; this chapter's account can help supply that lack.

This chapter proceeds as follows. Section 2.2 presents a general discussion of the indirect evolutionary approach to emergence of behavioral types in game theory and proposes an extension to situations where repeat players gain reputations of a general model already seen in the economics literature for the evolution of behavioral types in games with continuous action spaces. Section 2.3 discusses common features of political institutions with repeat participants that are well described by this general model. Section 2.4 applies the reputational extension of the evolutionary model to a particular game with a continuous action space. Section 2.5 elaborates on the connection to Kreps and Wilson (1982), and section 2.6 demonstrates application of the principles behind this evolutionary approach to a game with a discrete action space, where the previous results from the economics literature do not apply.

2.2 Theory

2.2.1 Background

2.2.1.1 Irrational Players

The rational-choice paradigm has attracted criticism from within political science and from outside precisely because the assumptions that define rational play are not supported by empirical observations of actual humans' behavior (Tversky and Kahneman, 1986), including their play in instantiations of game theoretic models from political sci-

ence (Ostrom, 1998). Tversky and Kahneman (1974) create a psychological paradigm for understanding judgment in terms of what they call cognitive heuristics and biases; a more quantitative description of how probability judgments are made under uncertainty is developed by Kahneman and Tversky (1979) under the name of prospect theory. One venture which the behavioral economics literature has engaged in relatively little, however, is the search for theoretical microfoundations of heuristics and biases; such microfoundations would be of use in predicting the ways in which humans will deviate from rationality under particular circumstances.

2.2.1.2 Modeling Evolution

A commonly suggested principle for solving games, one that is more fundamental than rational choice, is the evolution of players. If the payoff in a game represents or correlates with the likelihood of continuing to play the game, then players whose strategies yield higher payoffs will out-compete others whose strategies yield lower payoffs. In the long run, players whose strategies are not formed by rational choice can be expected to end up playing Nash equilibria on average due to selection in favor of these types. For biological systems that can be modeled as games, this selection is natural, occurring due to differential survival and production of offspring. In economic systems, by contrast, evolution can be expected as firms that realize higher profits more reliably avoid bankruptcy; in political systems, selection will operate on politicians' differing skills at getting elected, appointed, or retained in their offices.

The original evolutionary solution concept is the evolutionarily stable state of Maynard Smith and Price (1973). This concept is straightforwardly a refinement of Nash equilibrium; it is global and static. Young (1993) proposes a local but similarly static concept, the stochastically stable state. These static concepts ask which strategies outperform newly introduced mutant strategies given the new distribution of strategies after mutation, but do not explicitly describe the process of change. In contrast, dynamic evolutionary models set

up equations to represent the process by which individuals of different types enter or leave a pool of players as a function of payoffs, whose fixed points represent possible long-run populations. An example of the latter is the replicator equation (Gintis, 2009). Although evolutionary solution concepts are almost always refinements of Nash equilibrium, their predictions need not overlap; games with fixed points under the replicator dynamic may have no evolutionarily stable states and vice versa.

2.2.1.3 Evolution of Preferences

Besides explaining why players would play Nash equilibria, evolution as a microfoundation has been harnessed to explain why players assumed to play Nash equilibria would exhibit deviations from rationality. Guth and Yaari (1992) introduce the "indirect evolutionary approach" to motivating irrationality; this approach treats rationality and one or more kinds of irrationality as types subject to evolutionary pressure resulting from their performance in a repeated stage game. Players play equilibria of the stage game according to the type of preferences they are assigned, which may or may not be the same as the game's true payoffs; the latter determine the evolution of the pool of players.

Whether or not evolution will lead to the emergence of irrational players in general games depends on the model of evolution itself. Using static solution concepts, Ely and Yilankaya (2001), Ok and Vega-Redondo (2001), and Sethi and Somanathan (2001) show that non-material preferences, such as altruism, spite, or reciprocity, can be expected to emerge in general games only under knife-edge conditions where players perfectly know which types they face. By contrast, Heifetz, Shannon and Spiegel (2007*b*) adopt a dynamic evolutionary model and show that the limiting distribution of preferences should almost always include irrational preferences of some kinds, such as altruism or spite and optimism or pessimism, except when players get no information at all about each other's preferences. In each picture, long-term survival of preferences other than fitness maximization hinges on whether players know their opponents' quirks.

2.2.2 A generic evolutionary model

Formally, consider a continuous population of players repeatedly paired against each other randomly to play a two-player, symmetric game with a continuous pure strategy space $x^i, x^j \in [\underline{x}, \overline{x}] \subset \mathbb{R}$ for players i and j.[1] Let each player i be characterized by a fixed disposition $\tau^i \in [\underline{\tau}, \overline{\tau}]$. Denote by $\Pi^i(x^i, x_)$ the objective, material payoffs received by player i when she plays x^i and player j plays x^j . Suppose that players play equilibria of the game not according to their material payoffs Π but subjective, perceived payoffs $U^i(x^i, x^j, \tau^i)$, where $U^i(x^i, x^j, \tau^i) = \Pi^i(x^i, x^j) + \beta(x^i, x^j, \tau^i)$; that is, let $\beta(x^i, x^j, \tau^i)$ be a function relating a player's disposition to the amount by which they *misperceive* the material payoffs to any given action in the game. Denote by $\hat{x}^i(\tau^i, \tau^j)$ the action of player i in the equilibrium according to the players' subjective payoffs.

Now suppose that the distribution of dispositions evolves over time according to the *replicator dynamic* (Oechssler and Riedel, 2001), which states that the instantaneous rate of increase in the population fraction of any type of player is proportional to the difference between the average material payoff realized by players of that type and the average material payoff realized across the whole population at that time. Formally, given a distribution of dispositions $Q(\tau)$ (where $\int_{\underline{\tau}}^{\overline{\tau}} Q(\tau)d\tau = 1$), the expected payoff to a player of type τ^i is given by

$$E[\Pi(\tau^i)] = \int_{\underline{\tau}}^{\overline{\tau}} E[\Pi(\tau^i, \tau^j)]Q(\tau^j)d\tau^j) \tag{2.1}$$

and the instantaneous change in the proportion of the population made up of types within a given subset A of the space of dispositions is therefore

[1] Many politically relevant two-player games are not symmetric, but can easily be treated as symmetric by the fiction of a move by nature assigning individual players to roles in the game with probability one half each. Many politically relevant games have pure strategy spaces with dimension greater than one, but as long as the game contains a subgame in which the strategy space consists of an interval in $\mathbb{R}$, these results will continue to apply, albeit with increased complication.

$$\begin{aligned}\frac{\partial}{\partial\tau}Q(A) &= \int_A \left(E[\Pi(\tau^i)]) - \int_{\underline{\tau}}^{\overline{\tau}} E[\Pi(\tau^k)]Q(\tau^k)d\tau^k \right) Q(\tau^i)d\tau^i \\ &= \int_A \left(\int_{\underline{\tau}}^{\overline{\tau}} E[\Pi(\tau^i,\tau^j)]Q(\tau^j)d\tau^j - \left(\int_{\underline{\tau}}^{\overline{\tau}} \int_{\underline{\tau}}^{\overline{\tau}} E[\Pi(\tau^k,\tau^j)]Q(\tau^j)d\tau^j Q(\tau^k)d\tau^k \right) \right) Q(\tau^i)d\tau^i \end{aligned} \tag{2.2}$$

Further, fix some $\rho \in [0,1]$; let each player be commonly known to begin the game with perfect information about the opponent's disposition with probability $1-\rho$ and with no information other than the population distribution of dispositions $Q(\tau)$ with complementary probability ρ.[2] The value of $\hat{x}^i$ will then depend on both whether player i's type is disclosed to player j and vice versa. Thus, there is a function, which we can think of as a "types game", from players' dispositions, given the state of the population and the parameters of the game including ρ, straight to their expected material payoffs. Strategies in the underlying game are induced by subjective utility-maximizing play, but the evolutionary process still determines long-run strategies through the solution it places on the types game. Denote by ω the mean disposition in the population, and denote by $BR(\tau^j;\rho,\omega)$ a disposition for player i that is a "best response" in the types game, in the sense that this disposition maximizes expected material payoffs $E[\Pi^i]$, given ρ and ω, when facing an opponent of type τ^j. Then Theorem 4 in Heifetz, Shannon and Spiegel (2007*b*) establishes that the population of players converges over time to a monomorphic population in which all players have disposition

$$\hat{\tau} = BR(\hat{\tau};\rho,\hat{\tau}) \tag{2.3}$$

and that $\hat{\tau} \neq 0$ in general for $\rho < 1$, under the following conditions on the types game:

1. $BR(\tau;\rho,\omega)$ is single-valued everywhere;

[2] Heifetz, Shannon and Spiegel (2007*b*) treat disclosure as always mutual; either both players' dispositions are disclosed or both remain concealed. This symmetry is not required, however; the proof of their result does not depend on the game-theoretic meaning of ρ, but simply uses it as some parameter in the unit interval on which the game depends.

2. $E(\Pi^i)$ has the single crossing property in either (τ^i, τ^j) and $(\tau^i, -\omega)$ or $(\tau^i, -\tau^j)$ and (τ^i, ω); and

3. BR is Lipschitz continuous in τ and ω with constants k_τ and k_ω such that $k_\tau + k_\omega < 1$.

2.2.2.1 But why would dispositions ever be disclosed?

The indirect evolutionary approach as described above helps answer one question – why would those who play a game repeatedly converge over time time to a way of playing that is not payoff-maximizing? – but frames the answer in terms of another question it does not answer: why would players ever be presented with perfect information about each other's dispositions before play? Perfect information is treated as a knife-edge case in game theory, to be generalized if possible to some kind of imperfect information; the indirect evolutionary approach's main point depends on being in that knife-edge case, though, so agreement to that point should be supported by some reason to agree that the knife-edge case will obtain.

In the economics literature to date on this topic, relaxation of the assumption of perfect information is accomplished by assuming that players get perfect information with some probability strictly less than one and constant across all players and populations, and do not get to see opponents' types otherwise (Dekel, Ely and Yilankaya, 2007, e.g.). Further relaxations such as in Heifetz, Shannon and Spiegel (2007*b*) may include the random realization of a noisy, rather than precise, signal of the opponent's disposition. Other model features are not altered, including the random matching of players from a continuous population.

What kind of process, though, would indiscriminately disclose to players the exact dispositions of their randomly matched opponents even some of the time? That is, why should a knife edge be landed on with any positive probability, whether or not with probability one? Even if signals about opponents' dispositions are noisy, what is there in the world to be modeled that provides these signals at all? Below, I start from stylized facts

that characterize many political institutions to create a model of one such process for perfect disclosure of dispositions, a model hopefully more substantively believable than the state-of-the-art model of wholly random disclosure for no reason.

2.2.3 Extension: a generic model with reputation

In particular, I propose to model the perfect disclosure of a player's disposition to future opponents as a deterministic consequence of their past plays of the game. A player who, in some previous iteration of the game, has made some move that only someone with exactly her disposition would make in that situation will face opponents in the future who know her disposition exactly. I add this feature to the model of disposition evolution otherwise developed in Heifetz, Shannon and Spiegel (2007*b*); where they specified an arbitrary, fixed probability of disclosure, I allow the proportion of players whose disposition is disclosed to evolve over time, driven by the same selection process that drives evolution of dispositions themselves, and reach a long-run value. This long-run value will generically be non-zero. Thus, the knife-edge character of the assumption of perfect information at least a fraction of the time will greatly reduced.

To get there, consider the substantive foundations of the replicator dynamic. This dynamic for the evolution of player types may be justified by any number of qualitative descriptions of the reasons why the population of players changes over time. The replicator dynamic obtains as one model of learning by players in a permanently fixed pool who randomly switch to better strategies (or types) in proportion to those strategies' payoffs. The replicator dynamic also obtains as a model of reproduction, where the population grows as players generate offspring in proportion to their payoffs. Finally, and most interestingly for the study of political institutions, the replicator dynamic obtains as a model of survival, where the probability that a player returns to play the game again in the future is linear in the payoff the player receives in the current iteration. Since populations whose players all go away in the long run are of relatively little interest, the survivorship picture

must be supplemented with a picture of how new players enter the population to replace those who exit; the replicator dynamic continues to obtain under the assumption that the distribution of types among those entering the population at any time is equal to the distribution of types among those remaining in the population at this time. These are the situations to which this chapter applies most naturally. In section 2.3 further below, I will make the case that such a situation, where more successful players return to play again with higher probability and the replacements for those who leave have traits like those who remain, is a good picture of how many pools of interacting political players change over

Specifically, start from the model in section 2.2.2. Define $V : \mathbb{R} \to [0,1]$ as a linear, increasing function mapping $\Pi^i(x^i, x^j)$ to (a subset of) the unit interval. Let $V^i(x^i, x^j)$ be the probability that player i remains in the population after playing a game in which player i obtained material payoff $\Pi^i(x^i, x^j)$. Then the same replicator dynamic obtains for the transformed payoffs $V^i(x^i, x^j)$ as for the original payoffs $\Pi^i(x^i, x^j)$, up to a change in the speed of evolution.

Now, in addition to their disposition τ^i, endow each player i with a *status* $S_i \in 0, 1$. Let all players in the initial population at time $t = 0$ have status 0, and suppose that all new players who have never played before have status 0. When player i is involved in a game and makes a move that only a player of disposition τ^i would make in the situation player i faced, let player i's status be changed to 1, and let it remain equal to 1 until player i ceases to be part of the population. Suppose that when facing an opponent of status 0, a player gets no information about her opponent's type except for the distribution of types $Q(t)$, but when facing a player of status 1, she observes the opponent's disposition. That is, players get to know their opponents' dispositions if and only if some move their opponents made before pins it down, given the circumstances of the game the opponent was in. Since each player also knows her own history, it is common knowledge whether or not player i observes τ^j and vice versa.

From the conditions for the replicator dynamic (in particular, from the continuous-time approximation and the assumption that the distribution of entrant dispositions equals the distribution of dispositions among surviving players), it follows that the the proportion of players who have status 1 is equal across all player types at any given time. Then, in terms of the above model, the probability $1 - \rho$ that player i's type will be disclosed to player j is simply the proportion of players (at the current time) who have status 1.

At what rate are players with status 1 replaced by players with status 0 and vice versa? Since every player who leaves the game is replaced with a status-0 player, at any time t the rate at which status-1 players are replaced with status-0 players is simply $1 - \overline{V}_1(\omega; \rho)$, where $\overline{V}_1(\omega; \rho)$ is the population average V among status-1 players given ρ and average type ω. Likewise, some fraction φ of status-0 players disclose their dispositions by making a particular move at time t, so the rate at which status-0 players are replaced with those of status 1 is $\varphi\overline{V}_0(\omega)$ – the product of the probability of making a disclosing move and of remaining in the population to play again conditional on making a disclosing move.

Assume that φ is a constant, not dependent on τ^i, τ^j, ω, or ρ. In continuous action spaces this restriction will not be too demanding, but it rules out, for instance, games in which players with a range of dispositions pool on the lowest or highest move, as well as ones where all players prefer the same move for some values of ω or τ^j. As long as this restriction is satisfied, in place of the original equation 2.3 for the long-run disposition $\hat{\tau}$, depending on an arbitrarily fixed ρ, we may set up a system of two equations in the two unknown $\hat{\tau}$ and $\hat{\rho}$ to describe the steady state of our population:

$$\begin{aligned} \hat{\tau} &= BR(\hat{\tau}; \hat{\rho}, \hat{\tau}) \\ \hat{\rho} &= \hat{\rho}(1 - \varphi\overline{V}_0(\hat{\tau}; \hat{\rho}) + (1 - \hat{\rho})(1 - \overline{V}_1(\hat{\tau}; \hat{\rho})) \end{aligned} \tag{2.4}$$

The second equation expresses the condition that the concentration of status-0 players times the rate at which they are replaced by status-1 players equals the concentration of status-1 players times the rate at which they are replaced by status-0 players. Note, in

particular, that unless all players fail to survive the first time they make a make a disclosing move – that is, if $\overline{V}_0(\hat{\tau};\hat{\rho}) = 0$ – then $\hat{\rho}$ must also be less than one, even if $\rho = 1$ at time $t = 0$; for similar reasons, $\hat{\rho}$ must be greater than zero except in the trivial case where all players survive every time they play.

2.2.3.1 Existence, uniqueness, and convergence

Equation 2.4 describes the fixed point of a vector-valued function from the compact set $[\underline{\tau}, \overline{\tau}] \times [0,1] \subset \mathbb{R}^2$ to itself. To use this fixed point as a solution concept, it is desirable that the fixed point exist, that it be unique, and that the player pool actually converge to a monomorphic population with type $\hat{\tau}$ and undisclosed player fraction $\hat{\rho}$ given by the fixed point as $t \to \infty$. Continuity of the function for which the above is a fixed point implies that the fixed point exists. As long as $BR(\tau;\rho,\omega)$ is continuous in ρ as it has already been assumed in τ and ω, and as long as $\overline{V}_0$ and $\overline{V}_1$ are continuous in ω and ρ (since the function defining ρ is a linear combination of $\overline{V}_0$ and $\overline{V}_1$), continuity of the vector-valued function is ensured and the fixed point exists.

However, to guarantee uniqueness of the fixed point, and thus to guarantee that the long-run population of players is described by it no matter what the initial population, in the same manner would require assuming stronger conditions. For uniqueness to hold in general, it would have to at least be true that, for any given distribution of dispositions Q on date t, any two feasible values $\rho_t \neq \rho'_t$ of the fraction of players whose disposition is undiclosed on date t and some constant $c \in [0,1)$, the respective values ρ_{t+1} and ρ'_{t+1} of the fraction of players on the next date $t+1$ obey $|\rho_{t+1} - \rho'_{t+1}| \leq c|\rho_t - \rho'_t|$. Using the linear combination structure of the function mapping ρ on one date to ρ on the next would mean assuming this contraction-like condition on the average payoffs $\overline{V}_0$ and $\overline{V}_1$, a substantial constraint on the set of games under consideration far stronger than the super- or submodularity-like conditions already assumed by Heifetz, Shannon and Spiegel (2007*a*) – and would still provide only partial progress toward a guarantee of uniqueness.

Nonetheless, the conditions already assumed by Heifetz, Shannon and Spiegel (2007*a*), and the fact that $\rho \leq 1$ generally, suffice to guarantee the qualitative result of interest, that behavioral dispositions other than payoff maximization are not in general extinguished from the population in the long run when type information is provided through the reputation mechanism. If ρ converges to any value in the long run, then the original Heifetz, Shannon and Spiegel (2007*a*) results for fixed ρ must hold, yielding convergence to a monomorphic population with type $\hat{\tau}$. If ρ does not converge, then the distribution of types Q also must not converge to any monomorphic population: if all players in the population are of a single disposition, then players have perfect information about their opponents' dispositions whether those opponents have status 0 or 1, so $\overline{V}_0 = \overline{V}_1$ regardless of ρ itself; since φ is effectively 1 under perfect information, ρ_{t+1} will simply be equal to $(1 - \overline{V}_0)$ for any ρ_t. Lastly, if Q does not converge to a monomorphic population, it must not converge at all, under the condition already assumed that there is a well-defined best response in the types game. Thus, the only possible long-run outcome other than convergence to a monomorphic population described by $\hat{\tau}$ and $\hat{\rho}$ is a stable or chaotic limit cycle through the space of disposition distributions Q and fractions ρ of players with undisclosed distribution, in which case it will still be true that some players in the population on any date have preferences other than fitness maximization.

Further, while uniqueness of the fixed point is difficult to guarantee in general, it will be easy to check in any particular game, and checking at least local stability of the evolutionary dynamic around the monomorphic population defined by the fixed point should be possible, albeit tedious, in case the researcher is worried for some reason about this population having no basin of attraction. On this basis, it seems sensible to use the solution concept defined in this chapter even without general guarantees of its applicability – the qualitiative prediction of irrational preferences surviving in the long-run population holds in general and actual convergence to this solution from at least some starting positions can be verified if desired.

2.3 Stylized facts underlying this model

What makes this model work well at portraying the change over time in the distribution of dispositions among a pool of participants specifically in a political institution?

Political institutions are not anonymous. At least in many institutions of interest, there is a unique identifier – say, a name – associated with each person participating in the institution, and people who interact within the institution do at least get to know each other's names. Furthermore, because much of politics produces outputs that are matters of public record, participants' past actions within the institution can be learned about by knowing their names.

This assumption of non-anonymity may be tenuous for models of institutions like street-level bureaucracy, where records of caseworkers' dealings with one client may be unavailable or even prohibited to the next, or like courts issuing per curiam opinions whose author cannot be identified with certainty within the set of judges hearing the case, but this limit of low information about political actors and their behavior is not reached in many other institutions of interest. In some courts, opinion authors are known. Dollar amounts spent on today's election campaigns can be looked up by tomorrow's challengers. Pleas and sentences in criminal cases must be documented.

More successful players are more likely to continue playing. Elected politicians who diverge from constituents' policy preferences, or who deliver few goods and services, or who raise little money for campaigning are at greater risk than others of losing their future elections. Being worse at a political job should predict more likely leaving the job even for those not facing electoral incentives, however. Bureaucrats who do not meet the performance expectations of supervisors or political principals are at risk of being fired or of failing to rise within organizational hierarchies. Political actors who also have career or market options outside politics will be more likely to exercise these options and quit the

political institution the less they are personally benefiting from remaining in it.

Incoming players resemble existing players. Getting involved in politics happens to some people with greater probability than others. Bureaucrats with hiring power will seek out underlings who share their own goals and priorities for the office. Managers of past successful campaigns will seek to work with other probable winners who share the traits of their previous elected clients. Across many classes of institutions whose members are selected or elected, today's institutional newcomers will be selected at least in part for the extent to which they share the characteristics – the dispositions – of those already involved in the institution.

Incoming players are unknown. While new players on average look like surviving players in their political institutions, I do not assume that new players can be identified by name with any particular "parent" whose own disposition provides any information about theirs. This assumption might be unwarranted for relatively public figures; we may know that, say, the new attorney general is a political protege of the existing governor who shares both political views and personality characteristics with her mentor. However, the assumption that incoming players are unknown is the hardest case for the model; if every player's type were known to be identical to that of a specific previous player who recruited her, in the long run (indeed very quickly) all players' dispositions would necessarily be known.

Irrationality is a sticky trait. I assume that players are not learning to deviate from payoff maximization (or to exhibit other behavioral traits) more or less as they play the game.

Each of these common features of political institutions has a relatively sparse formal realization in the above evolutionary model. These features can be softened without compromising the main point that evolution of dispositions will generally favor some nonzero

value. Some anonymity may be introduced, for instance, by supposing that players may only randomly observe their opponents' dispositions with some probability less than one when playing status-1 opponents; this will simply insert a non-negative constant into both sides of the second equation. However, to make a simple, qualitative point about the kind of deviation from strictly rational behavior favored by the rules, as captured in a formal model, of an institution with at least some amount of these features, the additional mathematical trouble involved in adding generality to the model may not be warranted.

2.4 Example: a joint effort game

In this section, I will provide a demonstration of the above technology applied to a joint effort game, in which each player chooses an effort level and realizes a payoff that depends quadratically on both their own effort and (to some extent smaller in absolute value) the opponent's effort. Political examples of players in this kind of game might include executives in neighboring counties deciding on outlays for a shared infrastructure or transportation project, whose reelection prospects depend on the net value the project delivers to the county. We will suppose that players may be under- or over-confident about the linear component of the effect of their own action on their payoff. Heifetz, Shannon and Spiegel (2007*a*) use this game as an example to demonstrate the evolution of dispositions to a non-zero long-run value with fixed probability of random mutual disclosure; I will show the analogous evolution of dispositions and statuses.

Let each player $i \in \{1,2\}$ (where $j = 3 - i$ throughout) be allowed to select an action $x^i \in [\underline{x}, \overline{x}]$, where $\underline{x} \geq 0$. Let player i receive material payoffs $\Pi^i = (a - x^i - bx^j)x^i$, where $a > 0$ so that at least some non-zero action yields higher payoff than $x^i = 0$, and $b \in (-1,1)$, again so that the payoff-maximizing action is both bounded and nonzero. (Note that, for $b > 0$, the externalities each player's actions impose on the other are negative.) Let each player have a disposition $\tau^i \in (\underline{\tau}, \overline{\tau})$, where $-a < \underline{\tau} < 0 < \frac{a}{5} < \overline{\tau}$, and let each

player behave as if to maximize a subjective payoff $U^i = \Pi^i + x^i\tau^i$; the disposition τ^i is here characterized as the amount that each player enjoys the action itself, irrespective of how it pays off.

Now, suppose that a large pool of players participates in this game repeatedly, facing a randomly selected opponent each time, according to the evolutionary model in equation 2.2. Suppose, for simplicity, that $\Pi^i = V^i$, the probability of surviving to play again.[3] For the material payoff to equal a meaningful probability not greater than unity requires further constraints on the parameters; for $b \geq 0$, the relevant constraint is $a \leq 2$, while for $b < 0$ it must be true that $a - bx \leq 2$; which is to say that $\overline{x} \leq \frac{a-2}{b}$.

As shown in the original chapter, when both players' types are disclosed, they play a (subjective) Nash equilibrium in which $\hat{x}^i = \frac{2(a+\tau^i)-b(a+\tau^j)}{4-b^2}$, while when both players' types are not disclosed, they play a (subjective) Bayesian equilibrium in which each best responds to the opponent's expected move given the average type in the population, yielding $\hat{x}^i = \frac{1}{2}(a + \tau^i - b\frac{a+\omega}{2+b})$. Finally, in the case where player i's disposition is disclosed to player j but not the other way around, $\hat{x}^i = \frac{2(a+\tau^i)-b(a+\omega)}{4-b^2}$, while player j's move is the more complicated expression $\hat{x}^j = \frac{1}{2}(a + \tau^j - b\frac{2(a+\tau^i)-b(a+\omega)}{4-b^2}$. Since a given move in a given information environment always corresponds exactly to a particular type, $\varphi = 1$ exactly.

After substituting the above expressions into Π^i (that is, into V^i), equation 2.4 becomes

$$\begin{aligned} \hat{\tau} &= \frac{4a(2-b)b^2(1-\hat{\rho})}{4b^3(1-\hat{\rho}) + 32 + 2b^4\hat{\rho} - 8b^2} \\ \hat{\rho} &= \frac{4 - a + b^2 + ba\hat{\tau} + \hat{\tau}^2 + b(4+\hat{\tau}^2)}{(2+b)^2} \end{aligned} \tag{2.5}$$

We may observe before continuing that, for $b = 0$, the long-run disposition is greater

[3] Politically, of course, it is implausible that reelection odds for an official exactly equal the fraction of theoretically available returns they realize from a single joint project, but as long as these survival odds are at least monotonically increasing in those returns, exactly the same evolutionary dynamics will obtain, up to a change in the time scale of the evolutionary process and a monotonic, sign-preserving transformation of the long-run disposition; that is, results will be qualitatively similar regardless of exactly how important payoffs in this game are to political survival.

than zero. No matter whether players impose positive or negative externalities on the other, the disposition that is favored will be one that supposes one's own spending to be more efficacious than it actually is.

Substituting one equation into the other gives a cubic equation for $\hat{\tau}$ or $\hat{\rho}$ in terms of a and b. Of the three potential solutions, at the meaningful one it must be true that $\hat{\rho} \in [0,1] \subset \mathbb{R}$; there is at most one such solution at any point in the region of feasible parameters where $0 < a < 2$ and $-1 < b < 1$.

Figure 2.1 plots $\hat{\rho}$ with respect to a and b. Observe that $\hat{\rho} = 1$ for $a = 0$ and $\hat{\rho}$ is decreasing and concave down in both a and b. Note also that, for high a and low b, $\hat{\rho} < 0$, failing to represent a meaningful probability. In this region, it must be the case that $\hat{x}^i$ is greater than the upper bound on the maximum move set by $a - bx \leq 2$ in at least one of the possible information situations player i could face. What happens for these parameter values is not plotted.

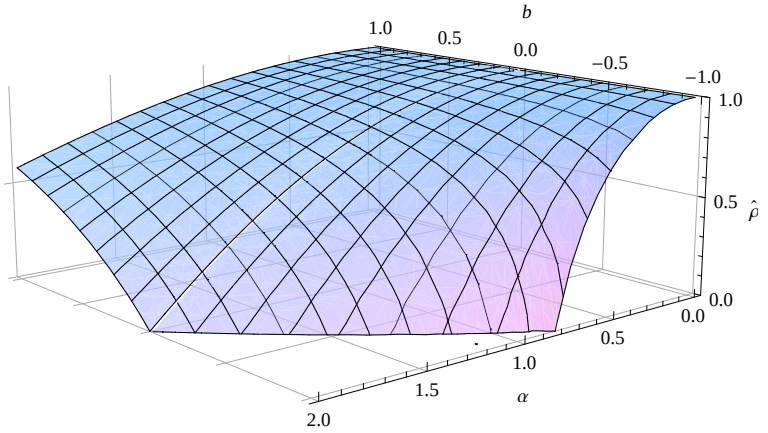

Figure 2.1: $\hat{\rho}$ vs. a and b

Figure 2.2 plots $\hat{\tau}$ with respect to a and b. (Due to numerical instability, the region on which $\hat{\tau}$ is not plotted is slightly larger than the region on which $\hat{\rho}$ fails to be meaningful.) Notice that, for $b > 0$, $\hat{\tau}$ is increasing in b, implying that, for fixed a, $\hat{\tau}$ is increasing in

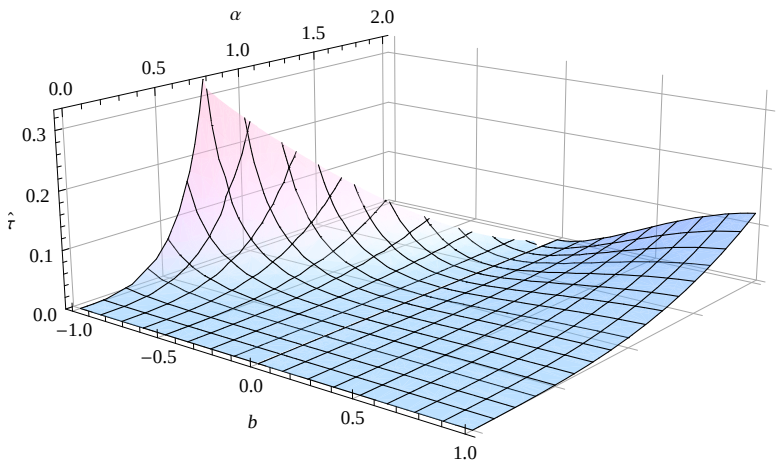

Figure 2.2: $\hat{\tau}$ vs. a and b. Note reversed orientation vs. Figure 2.1.

$\hat{\rho}$, even though $\hat{\tau}$ is decreasing in ρ for ρ free! Allowing the long-run disposition and the long-run probability of disclosure to coevolve produces, in part of the parameter space, results that would not be expected if the latter were left to be specified arbitrarily.

As mentioned above, the long-run fate of the population for the highest values of a and lowest b, where the constraint is binding that $\overline{x} \leq \frac{a-2}{b}$, is not investigated. Equation 2.5 will be replaced on these parameter space regions with a more complicated set of conditions on this region of the parameter space. Whether these parameters will lead to a monomorphic population with a nonzero disposition or to a stable limit cycle through the space of dispositions is not clear, but, unless the equlibrium moves and expected payoffs to playing them change discontinuously on this region, it will not be the case that bias disappears from the population for these extreme parameter values any more than for more moderate ones.

2.5 Discussion

2.5.1 The Medium and the Long Run

This chapter is not the first in the literature to predict behavior that diverges from single-period payoff maximization when players can gain a reputation for holding nonstandard preferences. Kreps and Wilson (1982) study a finitely repeated version of the chain store game of Selten (1978), in which a new firm decides whether or not to enter a market controlled by an incumbent monopolist and the incumbent then decides whether to acquiesce in the entry or fight a price war; the price war is the costliest option for both parties. Assuming that there is even a small probability of an incumbent with a strange preference for fighting, most incumbent firms will fight any entry that occurs until close to the end of the repeated game, because they profit in the long run by deterring future entries. Their account does not impose the substantive assumption that all players care about only today's payoffs rather than their expected lifetime utility, which this chapter's model does impose out of epistemological necessity; it is unclear what it even means to maximize an expected lifetime subjective payoffs when today's perceived payoffs may differ from the material rewards that determine the probability of playing again tomorrow.

However, where this chapter does not impose any particular assumption on what preference types its myopic players hold, instead insisting only that those types may vary among humans generally and that which humans are players will be evolutionarily determined over time by which ones perform best in the game, Kreps and Wilson (1982) do impose a substantive assumption on preference types: their result concerns only the situation where most players are of a fully rational preference type but a few have a bias toward toughness. They do not provide a reason that, in the long run, there should actually exist any incumbents who do intrinsically prefer to fight. In fact, a simple evolutionary argument shows that tough incumbents should not exist in the long run.

Suppose – perhaps not controversially – that firms in a market earn some kind of ma-

terial payoffs based on their performance in the finitely repeated chain store game. What does it mean for a new firm to face a small probability δ that the market incumbent is a tough type would rather fight than acquiesce in their market entry? Three apparent possibilities can be distinguished:

- No actual incumbents are of a tough type, but the challenger has an incorrect belief;
- Some incumbents receive higher material payoffs when they fight than when they acquiesce; or
- Some incumbents maximize a subjective payoff function that is not always increasing in material payoffs.

The language used in Kreps and Wilson (1982) makes clear that the first is not the story they believe, and, if it were, the bounded rationality thus attributed to the challenger firm(s) would itself ask for some kind of explanation. The second is substantively implausible in the setting to which the game is applied as a model; how a firm can make money rather than lose it by engaging in a price war is not clear. Perhaps it is to be taken for granted that they can do so, but if the reader suspends disbelief that far, there is no paradox for the reputational effect to explain. If they are neither figments of entrants' imaginations nor money sources of an unbelievable (or uninteresting) kind, tough firms must be behavioral types maximizing something other than profits.

If the tough incumbents are realizing the same *material* rewards as ordinary ones, then, by the end of the finitely repeated game, the expected total payoff earned by an ordinary player must be higher than that earned by a tough one. In any repetition of the game where the ordinary player acquiesces, their expected lifetime payoff from doing so must be at least as high as that expected from fighting as the tough player would. In at least one situation, the last repetition, the expected lifetime payoff from acquiescing is strictly higher than that from fighting as the tough player would, and the ordinary incumbent

does acquiesce. Lastly, on any turn on which the ordinary player fights or the challenger does not enter, the ordinary player earns exactly what the tough one would.

Now suppose that the finitely repeated game is the stage game of an evolutionary process with non-overlapping generations: after the last repetition, existing incumbents die out and new ones take their place, with the proportions of tough and ordinary in the next generation determined by the replicator dynamic. Since the tough incumbents achieve lower lifetime utility than the ordinary ones on average, this evolutionary model would predict that there will be fewer tough incumbents in each successive generation than the previous one, no matter how many there were to start. Tough players in this model must eventually go away.[4]

If the tough players go away, though, then so does the tough behavior by ordinary – rational – players. In the sequential equilibrium of the finitely repeated chain store game, for a given number of total repetitions, the expected number of repetitions in which the ordinary incumbent acquiesces is increasing in the initial probability δ that the incumbent is tough.[5] With each successive generation, then, the ordinary incumbents expect to fight on fewer opportunities. Kreps and Wilson (1982) thus provide a medium- but not long-run explanation for apparently tough action on the part of instrumentally rational actors.

[4]Strictly speaking, it is not yet ruled out that, due to a discontinuity at some $\delta' > 0$, the advantage experienced by the ordinary incumbent over the tough one approaches zero as δ approaches δ' from above, even though the advantage experienced by the ordinary incumbent is non-zero at exactly δ', in which case δ could approach δ' rather than 0 in the long run. However, such a discontinuity, while not yet ruled out, is not suspected. Furthermore, if the assumed evolutionary dynamic included a finite but arbitrarily small amount of random noise, the proportion of tough incumbents could not remain above δ' forever with positive probability. Put another way, the ordinary preferences are evolutionarily stable for all δ (albeit only weakly for $\delta = 1$), while tough preferences are not even weakly stable for any $\delta < 1$.

[5]To recap the sequential equilibrium: let the number of time periods n be N, with period $n = 1$ being the last and $n = N$ the first, and let the probability at time n that the incumbent is tough be p_n. For $p_n > b^n$, where $b < 1$ is the payoff to the challenger from entering when the incumbent acquiesces, the defense will enter with probability 0, and p_{n-1} will be set to $\max(b^{n-1}, p_n)$, while for $p_n = b^n$, the entrant will enter with probability $1/a$ where a is the payoff to the incumbent from non-entry, and the ordinary incumbent will fight with probability not dependent on δ. If the incumbent ever acquiesces, they make it clear that they are not tough, so from then on the challenger will always enter and the incumbent will always acquiesce. For $\delta > b^N$, there will always be some turns of non-entry followed by some turns on which the challenger and the ordinary incumbent randomize with probabilities independent of δ. The number of initial turns of non-entry, and hence the expected number of turns until an acquiescence occurs, is increasing in δ, since the value of n such that $b^n = \delta$ is decreasing in δ for $b < 1$.

By contrast, a long-run population of behavioral players maximizing single-period subjective rewards according this chapter's model can be at least neutrally stable with regard to invasion by farsighted, zero-bias players. Suppose that on some date t, such an invader were one of the new players in a large population of behavioral players repeatedly playing the joint-effort game above, with $a = 1$ and $b = -0.75$. Under these parameters, $\hat{\tau} \approx 0.10247$ and players of type $\hat{\tau}$ expect a material payoff of $\Pi = 1 - \rho \approx 0.59872$ each turn, or $\frac{1-\rho}{1-(1-\rho)} \approx 1.49203$ over a lifetime. If the invader's move today sets the other players' expectations about the invader's own (hypothetical) disposition, just as a behavioral player's first move would set others' expectations about their true disposition, what move today will maximize the invader's lifetime payoff?

In the long run, new players will all have type $\hat{\tau}$, so, if the invader is not distinguishable from other players with no history, the invader's opponent on date t will provide the same effort level as all other players. The invader can then get a higher payoff on date t than the behavioral players by providing only a level of effort that maximizes today's payoff while taking advantage of their opponent's generous effort level, at the cost of being treated as a zero-bias player on future turns. Is such a disruptive move worthwhile? In fact, it is not. The invader would get approximately 0.60154 on the first turn by playing a payoff-maximizing move and 0.59169 on future turns when opponents treat the invader as a zero-bias player rather than as one of them, for an expected lifetime payoff of approximately 1.47324. The invader does not have a move that yields a higher expected lifetime payoff than that earned by the behavioral players, and thus no way to (strictly) displace those behavioral players from the population. While a full dynamic analysis of the coevolution of farsighted and behavioral players would be intractable, the possibility for robustness to invasion suggested by the preceding static treatment gives this chapter's model a claim on being a long-run explanation for the prevalence of behavior that does not maximize short-term payoffs.

2.6 Discrete action spaces

Many institutions, treated as the rules of a game, are best modeled as affording political players only finitely many possible moves. A legislator can vote yea or nay on a given bill, for instance, but cannot in most cases cannot vote for an arbitrary convex combination of yea and nay. In these discrete evolutionary games, future opponents may never get enough information to fully pin down their opponents' dispositions just by recalling those opponents' previous moves. The general prediction of Heifetz, Shannon and Spiegel (2007*a*) of a long-run population of players with identical nonzero disposition thus does not extend fully to discrete games. Particular discrete games may nonetheless generate long-run populations of players with nonzero dispositions, and the populations generated when opponents' dispositions are only known when they are attested by prior moves may be qualitatively dissimilar in a way not seen for continuous games. This section presents those results for one particular discrete game.

Consider two legislators each interested in securing the last seat on the same committee, or interested in being named its chair. The committee position would serve to enhance either legislator's reelection prospects at the end of the term. Each legislator can either go public with a demand (perhaps costly) for their party leader to grant them the committee position or do nothing; if only one legislator demands the position, the party leader will grant it to them, but if both demand it, the leader will choose a third compromise candidate and punish both legislators for undermining party unity. If such conflicts of interest affect some legislators from the party with each new term, they will exert an evolutionary pressure favoring some kinds of legislator personalities over others; what types will emerge in the long run? Will they be more or less inclined to make demands than ones who simply maximize their own reelection probabilities?

The legislators described above are playing chicken, one of the canonical 2×2 games. Chicken, elsewhere called the Hawk-Dove Game[6], has the following normal form:

[6]Smith (1982) uses the term Hawk-Dove Game to refer to a slightly different set of games, imposing the

	D	H
D	(a, a)	$(b, 1)$
H	$(1, b)$	$(0, 0)$

where $1 > a > b > 0$. (H, D) and (D, H) are its two pure-strategy equilibria, and there is a third equilibrium in mixed strategies in which each player plays D with probability $\frac{b}{b-a+1}$.

The players in chicken might well be understood as politicians other than legislators competing for committee positions. Persons ambitious to enter a cabinet or become a party leader would fit well, as would an executive and a legislative leader promoting divergent national budgets, of which at least one must pass to avert a government shutdown. Chicken can also be seen as the simplest possible example of crisis bargaining models used in international relations, where war ensues if neither state accedes to the other's demands for an indivisible resource. In short, many interactions in which politicians repeatedly participate have a chicken flavor.

2.6.1 Chicken as an Indirect Evolutionary Game

To consider chicken as a behavioral game, consider players who mis-percieve the material value of the "sucker's payoff" b. If each player i undervalues the payoff to (D, H) by an amount τ^i , the payoff matrix of the game with behavioral preferences is given by

	D	H
D	(a, a)	$(b - \tau^i, 1)$
H	$(1, b - \tau^j)$	$(0, 0)$

It is sensible to assume the disposition $\tau^i \in [b - a, b]$; for dispositions outside this range, the player will not perceive the game to be chicken.

restriction (in this chapter's notation) that $a = \frac{b+1}{2}$ and relaxing the restriction that $b > 0$. This set of games has the same equilibria as the ones covered in this chapter.

2.6.2 Perfect Information

As a foil for later results, consider how dispositions would evolve if players did have perfect information about each other's dispositions. Since chicken violates Assumption A in Heifetz, Shannon and Spiegel (2007*a*) of a single pure-strategy equilibrium, it is necessary to specify on which equilibrium the evolutionary dynamics apply. One possibility is that players simply use each others' dispositions like sunspots, agreeing on some function that assigns one player the move H and the other D given their respective dispositions. Uncountably many rules for making this assignment could be imagined, and for any arbitrary set of dispositions there is an assignment rule that gives the long-term distribution of dispositions support on precisely that set. Obvious examples include a rule that requires the player with higher τ to play H, leading to a monomorphic population with maximum τ, or the reverse; a less obvious example might require the player with higher τ to play H if the sum of the most significant unequal digits of both players' dispositions is even and D otherwise, which would perfectly preserve an initial uniform distribution of dispositions into the long run. However, in these trivial outcomes the fact that τ^i is a bias away from objectivity in preferences does no work.

A less trivial result obtains if players do not use their dispositions to break the symmetry between them but instead play the mixed-strategy equilibrium of the stage game. In this equilibrium, each player plays D with probability $\hat{x}^i$ given by

$$\hat{x}^i = \frac{b - \tau^j}{1 - a + b - \tau^j} \tag{2.6}$$

As explained by Tsebelis (1990), each player's equilibrium mixed strategy is invariant in their own type, but tracks the mixing probability that would make the opponent indifferent given the opponent's type. The more each player is dispositionally inclined to like playing H, the more their opponent can be expected to play H against them. Since each player is always materially better off the more the opponent plays D, the most materially

profitable disposition is the minimum disposition $\tau^i = b - a$, regardless of the opponent's disposition. The replicator dynamics on the mixed-strategy equilibrium of chicken will thus converge to a monomorphic population of players with disposition $\hat{\tau} = b - a$ if players always know each other's dispositions perfectly.

This is a cooperation result; players minimize the frequency with which they play the aggressive pure strategy H. It may also be seen as unrealistic even given the assumption that players perfectly know each other's dispositions. In the mixed strategy equilibrium, players in the long-run population are all earning a in each interaction, but, in the less obvious example equilibrium above, players all average $\frac{b+1}{2}$. If $\frac{b+1}{2} > a$ (for instance, if there is any possibility of the resource going to waste if neither player demands it), then for all populations not too different from the mixed-strategy equilibrium's long-run population, the less obvious example equilibrium is Pareto superior to the one in mixed strategies; another even more bizarre equilibrium could necessarily be devised sufficiently similar to the less obvious example as to also be Pareto superior to the mixed-strategy equilibrium, but just different enough to evolutionarily favor a set of types other than $\{b - a\}$. While a complete treatment of the coevolution of equilibria and dispositions is beyond the scope of this chapter, it is generally believable that Pareto inferior equilibria are subject to long-run extinction. Since any equilibrium that uses disposition only to assign H and D roles to players yields an average payoff of $\frac{b+1}{2}$ to any population, however, equilibria in this class are not Pareto ordered; none is obviously more or less likely to emerge from a given initial condition.

By allowing this uncountable set of equilibria in which disposition matters, but only trivially, the perfect-information assumption greatly hinders any attempt to predict how that disposition will evolve. By contrast, the remainder of this section will show that predictive capacity is improved by treating the origin of players' knowledge about opponents' types as arising only from those opponents' past play. The more realistic treatment eliminates all the purely trivial equilibria and forces the disposition to do meaningful work in

at least some contests.

2.6.3 When disposition is disclosed only through prior moves

Now consider a chicken game being played by players who learn about their opponents' dispositions only by seeing what those opponents have done in the past. If players' dispositions initially follow a continuous distribution on an interval of the real line, there is zero probability that any player's disposition will be perfectly disclosed after finitely many plays of the game. The most information that yesterday's move can provide about the disposition of today's opponent is whether it is an element of the set of dispositions that would have played H or of the set that would have played D (with positive probability) in yesterday's game; at most one of these sets can be a singleton, and, if it is, there is zero probability of today's opponent having a disposition within that set.

For this reason, trivial equilibria where players always use each other's dispositions merely to assign one to play H and the other D do not arise when dispositions are disclosed only through previous moves. The most players could do in this regard would instead be to adopt a function mapping the expected values of their disposition and of their opponent's to the pure strategy set. However, it must always be the case that a nonzero fraction of games played on any date t are played between opponents with equal expected disposition (if nothing else, this will be true of any pair of opponents neither of whom has played a game before, since the dispositions of new players on any date are *iid* under the conditions of section 2.3). In these symmetric-information matchups, any difference in fitness to players of different dispositions must be non-trivially related to the actual preference divergence those dispositions signify.

If two opponents do not perfectly know each other's dispositions, they also do not individually play mixed strategies that make each other indifferent between H and D. Rather, Tsebelis (1990) shows that, if every player in one group is randomly matched against every player in another group, and individual dispositions within each group are *iid* on some

interval, then each group will play a mixed strategy on the group level, rather than the individual level. If the mixed strategy that would make player i indifferent in expectation, given player j's information, is to play D with probability β, then player j will play D if player j has a disposition in the lowest β of the distribution of possible dispositions for player j given player i's information, and H otherwise. In these group mixing strategies, unlike in the mixed strategy with perfectly known dispositions, one's own disposition does affect one's move directly, while one's expected disposition on any date determines the average move of one's opponent.

No matter what equilibrium of the repeated game players engage in, any two opponents who see identical distributions over each other's dispositions must use a group mixing strategy. Thus, if on date $t = 0$, all players are brand-new and their distributions are *iid* on some interval, then on date $t = 1$ each player will be identifiable with exactly one of three distributions of possible dispositions. Those who played H (D) and survived will be known to be iid according to the distribution of types at time $t = 0$ truncated to have support only above (below) some threshold diposition; players who enter the population to replace those who did not survive will be *iid* on the whole interval, but not according to the same distribution from the previous period unless those who played H and D survived at identical rates.

From each date to the next, there will always be more distinct classes of players whose dispositions are *iid* according to different distributions on different subsets of the original interval. This proliferation of information situations in which any pair of opponents could find themselves will make a dynamic solution for the long-run distribution of dispositions intractable. Rather, this chapter will proceed by identifying which, if any, disposition values are candidates to take over the whole population by virtue of satisfying a static selection criterion – that of being an evolutionarily stable strategy (Smith, 1982).

Static analysis of the prospects for evolutionary emergence of a single disposition will be accelerated by two consequences of the conditions in section 2.3. First, if all players in

the population have the same disposition, then their opponents all have perfect information about their dispositions. Second, if players' dispositions do not change while they remain in the population, then any invaders in the sense used to define the ESS must not have played any games before; thus, even after invasion, players still have perfect information about their opponents' dispositions if their opponents have any history of play to observe.

Let the pre-existing population (before invasion) have disposition τ and play the mixed strategy $\hat{x}$ in equation 2.6 against each other. In the steady state, the fraction δ of those players who are of age greater than 1 will be given by their average fitness from playing this equilibrium:

$$\delta = \hat{x}(\hat{x}a + (1-\hat{x})b) + (1-\hat{x})\hat{x} \tag{2.7}$$

Now suppose that a small number of players with disposition $\tau' \neq \tau$ invade the population, so that the total fraction of the new population composed of invaders is $\epsilon \ll 1$. As above, all invaders have age 1 and players whose opponents have age 1 cannot tell whether their opponents are invaders or not. In any equilibrium between non-invaders with age greater than 1 ("adults"), non-invaders with age 1 ("children"), and invaders, adults will still mix when paired against each other, and age-1 players will play a group mixing strategy when paired against each other: children will mix, with probability approaching $\hat{x}$ as $\epsilon \to 0$, and invaders will play D if $\tau' < \tau$ and H otherwise. Thus, on the first turn after invasion, there are three equilibria obviously worth considering:

- one in which adults and age-1 players play a group mixing equilibrium against each other;
- a "focal" equilibrium in which each player plays H if the expected value of their disposition is higher than their opponent's; and
- a "polite" equilibrium in which each player plays D if the expected value of their

disposition is higher than their opponent's.

2.6.3.1 Group Mixing against Adults

In the first, adults and children all earn δ (in the limit as $\epsilon \to 0$), while invaders earn

$$V^I = \begin{cases} \hat{x}a + (1-\hat{x})b, & \tau' > \tau \\ \hat{x}, & \tau' < \tau \end{cases} \tag{2.8}$$

Since $\delta \to \hat{x}$ as $\tau \to 0$, $\delta > (\geq) V^I \Leftrightarrow |\tau| < (\leq)|\tau'|$. Thus, if adults and age-1 players use a group mixing strategy against each other, the only disposition that is a candidate to be an ESS is $\tau = 0$.

2.6.3.2 The Focal Equilibrium

The expected value of age-1 players' disposition is higher than that of adults if $\tau' > \tau$ and lower otherwise. Thus, for $\tau' > \tau$, fitness payoffs in the focal equilibrium are given by

$$\begin{aligned} V^A &= \delta^2 + (1-\delta)b \\ V^C &= \delta + (1-\delta)\delta \\ V^I &= \delta + (1-\delta)\hat{x} \end{aligned} \tag{2.9}$$

The existing population averages greater fitness payoffs than invaders with higher dispositions if τ is greater than the real root of the cubic equation

$$\begin{aligned} 0 = \quad & b - ab - a^2b + a^3b + b^2 + 2ab^2 - 3a^2b^2 - b^3 + 3ab^3 - b^4 \\ + \tau \quad & \times(-3 + 7a - 5a^2 + a^3 - 4b + ab + 2a^2b + a^3b + b^2 - 5ab^2 - 2a^2b^2 + 2b^3 + ab^3) \\ + \tau^2 \quad & \times(2 - 2a^2 + b + ab + 4a^2b - b^2 - 2ab^2) \\ + \tau^3 \quad & \times(-1 + a - 2a^2 + ab) \end{aligned} \tag{2.10}$$

On the other hand, for $\tau' < \tau$, players get fitness payoffs

$$\begin{aligned} V^A &= \delta^2 + (1-\delta) \\ V^C &= \delta b + (1-\delta)\delta \\ V^I &= \delta b + (1-\delta)\,(\hat{x}a + (1-\hat{x})b) \end{aligned} \tag{2.11}$$

If τ is less than the first root of the sextic equation in Appendix 2.A, then the existing population averages greater fitness payoffs than invaders with lower dispositions. The sextic and cubic surfaces do not coincide; dispositions above the cubic surface and below the sextic surface meet the letter of the definition of an ESS, while all others can be invaded. For any given a and b, which disposition values does the ESS solution concept predict may take over the whole population?

It can be seen algebraically that the real root of equation 2.10 is zero for $b = 0$ and increasing in a and b. Thus, only dispositions greater than 0 are candidates to take over the population in the evolutionary long run. If all players converge on any disposition, they will converge on one that enjoys playing H more than if they maximized fitness payoffs, not the opposite as in the mixed-strategy equilibrium of the perfect-information case. The fact that, in group mixing, players' own types influence their moves reverses the prediction made under conditions where only opponents' types determined players' equilibrium mixed strategies.

Since the sextic surface cannot be solved algebraically, it is plotted numerically in orange, along with the cubic surface in blue, in figure 2.3, and 2.4 plots the sign of the difference between the surfaces. For low a and b, the sextic surface lies above the cubic one; for these values, a narrow range of positive dispositions can fight off invasion by players with either higher or lower dispositions. For high a and b, however, the two surfaces cross, and the upper bound is below the lower one. Thus, in these cases there is no possibility of one disposition taking over; the long-run population is polymorphic and includes at least some players with nonzero disposition, although no further prediction is made

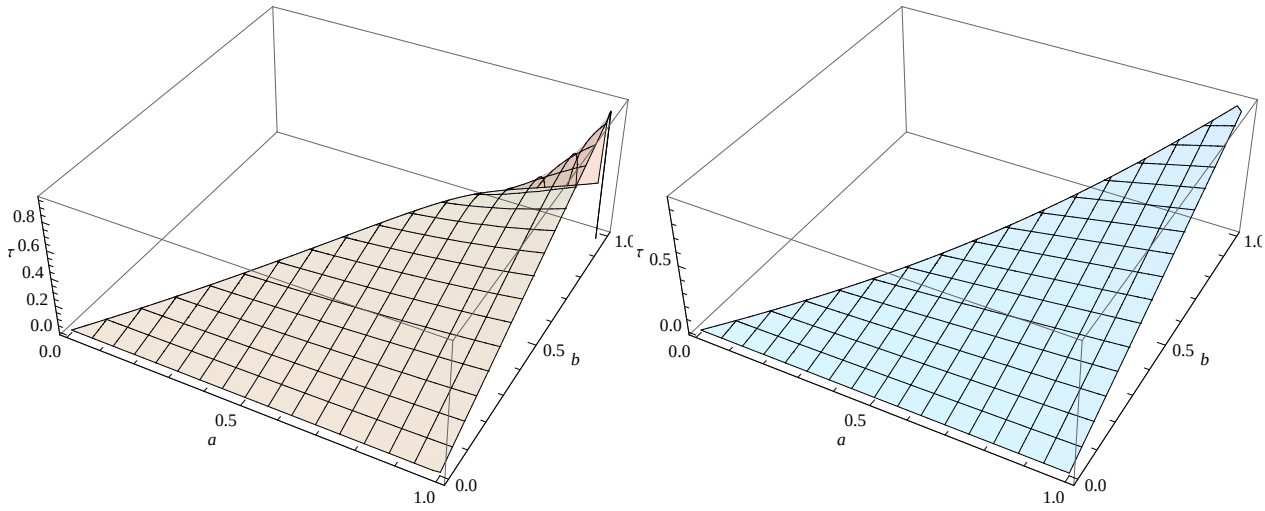

Figure 2.3: Upper (cyan) and lower (orange) bounds on $\hat{\tau}$ vs. a and b.

about which ones.

2.6.3.3 The Polite Equilibrium

In the third possible equilibrium, fitness payoffs are exactly dual to those in the second! For $\tau' > \tau$, players get fitness payoffs equal to those in the focal equilibrium with $\tau' < \tau$ and vice versa. The dispositions that are ESS in both the polite and focal equilibria are exactly the same – strictly greater than 0 for low a and b and an empty set elsewhere. Unlike in the perfect information case, the selection of an arbitrary rule for mapping pairs of dispositions to pure strategies does not determine which dispositions emerge in evolutionary competition if players get information on their opponents' dispositions only from opponents' past play, nor does the infinite possible variety of such mappings prevent making clear predictions about which dispositions do so. In fact, there are still uncountably many ways to map the dispositions of invaders and non-invaders to moves, but virtually all will still yield fitness payoffs given by equation 2.9 for some invader dispositions and by equation 2.11 for others. The prediction from the focal equilibrium is robust to equilibrium selection: for high enough payoffs to outcomes other than (H, H), the long-run population is diverse in disposition; otherwise, the game will select for aggressive play-

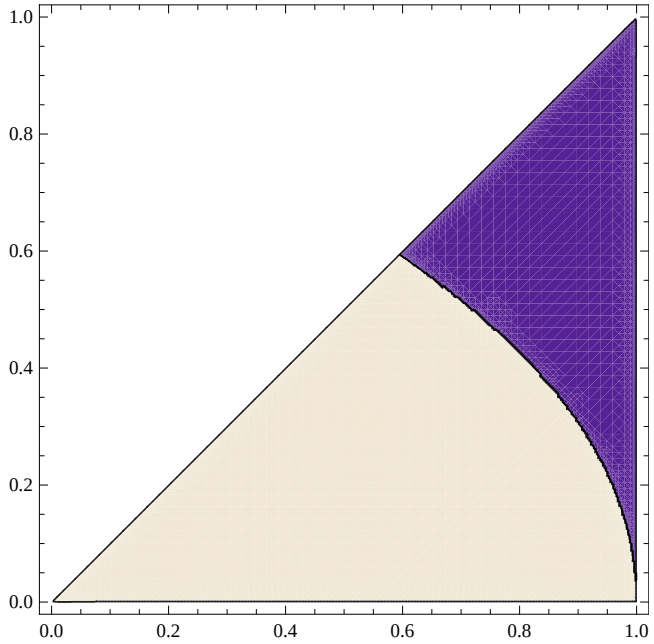

Figure 2.4: Sign of difference between upper and lower bounds on $\hat{\tau}$ vs. a and b. The lower bound is above the upper bound on the purple region and vice versa on the cream region.

ers.

2.7 Conclusion

Political institutions may select, by disproportionately preserving in office, for politicians who defy the strong versions of common rationality assumptions of game theory by maximizing payoff functions other than their objective probabilities of remaining in office (or who appear to outside observers to defy them). Modeling which non-fitness dispositions do in fact emerge can help bridge the gap between formal theory and the empirical behavioral studies highlighting real players' cognitive biases away from such rationality conditions. By definition, however, it must be opponents' responses to players' dispositions that end up doing the work of raising the player's realized fitness, so a model of this type should account for how players find out their opponents' dispositions. An obvious and realistic choice is to suppose that opponents' dispositions can be inferred from the moves they played in previous instances of the game. When the action space is continuous and

best responses vary continuously with opponents' dispositions, existing theory guarantees that there is a unique disposition that survives in equilibrium for any probability of facing an opponent with perfectly known disposition, and this theory extends naturally to guarantee that there is a unique long-run proportion of players whose disposition is known by virtue of having played once before. Existing theory does not cover the cases of discrete action spaces or multiple equilibria, but the assumption of preference disclosure through past play rather than perfect information can provide greatly improved predictive power in these situations as well.

2.A The Sextic Surface

The sextic surface in section 2.6.3.2 is given by

$$
\begin{aligned}
0 = \quad & -b+5ab-10a^2b+10a^3b-5a^4b+a^5b-3b^2+12ab^2-18a^2b^2+12a^3b^2-3a^4b^2 \\
& -3b^3+8ab^3-6a^2b^3+a^4b^3-4ab^4+8a^2b^4-4a^3b^4+3b^5-9ab^5+6a^2b^5+3b^6 \\
+\tau\times\ & (2-11a+25a^2-30a^3+20a^4-7a^5+a^6+8b-34ab+56a^2b-44a^3b+16a^4b \\
& -2a^5b-4ab^6+b^7+13b^2-40ab^2+42a^2b^2-16a^3b^2+a^4b^2+7b^3-7ab^3-5a^2b^3 \\
& +3a^3b^3+2a^4b^3-7b^4+23ab^4-10a^2b^4-6a^3b^4-11b^5+11ab^5+6a^2b^5-4b^6-2ab^6) \\
+\tau^2\times\ & (-5+22a-38a^2+32a^3-13a^4+2a^5-18b+61ab-76a^2b+42a^3b \\
& -10a^4b+a^5b-19b^2+37ab^2-21a^2b^2+7a^3b^2-4a^4b^2+2b^3-15ab^3 \\
& -5a^2b^3+17a^3b^3+a^4b^3+14b^4-22a^2b^4-2a^3b^4+6b^5+8ab^5+a^2b^5) \\
+\tau^3\times\ & (8-29a+40a^2-26a^3+8a^4-a^5+17b-37ab+25a^2b-7a^3b+2a^4b+7b^2-7ab^2 \\
& +21a^2b^2-19a^3b^2-2a^4b^2-6b^3-10ab^3+28a^2b^3+8a^3b^3-4b^4-12ab^4-4a^2b^4) \\
+\tau^4\times\ & (-5+11a-7a^2+a^3-7b+12ab-17a^2b+11a^3b+a^4b-b^2+10ab^2 \\
& -12a^2b^2-12a^3b^2+b^3+8ab^3+6a^2b^3) \\
+\tau^5\times\ & (2-4a+5a^2-3a^3+b-ab-2a^2b+8a^3b-2ab^2-4a^2b^2) \\
+\tau^6\times\ & (-a+2a^2-2a^3+a^2b)
\end{aligned}
\tag{2.12}
$$

Chapter 3

Ignorance is Strength: Evolution of Optimism Bias in a Plea Bargaining Game

3.1 Introduction

Of the millions of Americans who see the inside of a cell after a criminal conviction, only a small fraction see a jury inside a courtroom first. The majority of cases instead end with a negotiated guilty plea to charges that may fall short of those for which prosecution is initiated. For reducing courts' workload as much as an order of magnitude, the plea bargain system has often been praised as efficient. Plea bargaining as an institution, however, transfers power to determine penalties from judges and juries to attorneys, and mostly to the prosecutors who hold most of the cards in plea negotiations. Unlike judges and juries, prosecutors are representatives of only one side in an adversarial trial, and have incentives to "win" visible convictions in addition to, or even to the exclusion of, incentives to seek justice by separating the guilty from the innocent.

In this chapter I argue that, under such incentives, prosecutors who ignore at least some

information suggesting the defendant's innocence can outperform those who account rationally for such information in the plea offers they make. I present a game-theoretic depiction of plea negotiations at trial, in which the prosecutor and defense alternately negotiate over the ultimate sentence and fight in court. In the model, higher payoffs can be realized by prosecutors who fall short of Bayesian updating in the face of courtroom setbacks. This occurs because defendants with a low chance of winning at trial, such as guilty defendants, have a harder time bluffing and pretending to be hard-to-convict innocent defendants when the prosecutor does not fully update. The advantage to biased prosecutors will obtain precisely when prosecutors' choices are least clear: when the prosecutor's investigation is imperfect at revealing innocent defendants, and when innocent defendants are not much harder to convict than guilty ones.

Evolution of the population of prosecutors will occur as lower-performing types quit the job more frequently and are hired less frequently than higher-performing ones. Over time, career considerations will lead to such higher-performing attorneys predominating in prosecutors' offices. In the long run, therefore, we may observe most prosecutors to be biased toward optimism in regards to their ability to win trials, remaining unduly confident in the face of new revelations that the defendant will be hard to convict due to innocence. Optimism bias is an empirically established characteristic of human judgement, but prosecutors who get ahead in their jobs because of it it may be expected to disproportionately possess this trait.

Deviation from rational learning toward optimism bias among prosecutors will alter the pattern of plea bargains ultimately struck. In this model, optimism-biased prosecutors can be expected to more frequently induce guilty pleas before trial from guilty defendants, although their initial plea offers may be more lenient. Meanwhile, a biased prosecutor risks spending more time in court when facing innocent defendants, but sentences them to fewer overall years when this happens. A public that cares more about punishing the guilty and freeing the innocent than about the workload of the courts will prefer that its

prosecutors be biased.

This chapter proceeds as follows. In section 3.2 I discuss previous game-theoretic work on which this chapter builds. The game itself is presented and equilibria are found in section 3.3. In section 3.5, I proceed from equilibrium analysis to evolutionary results on the prevalence of biased palyers. Section 3.6 contains discussion of the game as a model of plea bargaining and of other political situations, and of the welfare implications of optimism bias in prosecutors. Finally, I present concluding remarks in section 3.7.

3.2 Background

3.2.1 Optimism Bias

As summarized in chapter 2, behavioral economists such as Tversky and Kahneman (1986) and other social scientists, including political scientists, have argued that microeconomic models, including game theoretic models, can benefit from the realistic relaxation of assumptions that players are perfectly rational to allow for the kinds of cognitive heuristics and biases humans consistently exhibit in lab settings.

The heuristics and biases approach has been applied in social psychology to the context of negotiation; de Dreu et al. (2007) highlight two such principles specific to the study of bargaining. First, *naive realism* is the assumption that one sees the world as it is, without accounting for one's own cognitive biases. Naive realism is shown to promote optimistic overconfidence (Kahneman and Tversky, 1995), described by Ross and Ward (1995) as subjects "believing that time is on their side, and that complete, unilateral victory is just around the corner", and also to drive a search for disproportionately confirmatory information (Rubin, Pruitt and Kim, 1994). These effects can add up to make expectations of other parties' aggressiveness in bargaining situations self-fulfilling (Diekmann, Tenbrunsel and Galinsky, 2003). Second, their *self-threat principle* states "that humans [...] develop, maintatin, and protect a positive self-concept and that evaluations of the self are

positively biased". Humans exhibit bias against judgments that would injure their self-concepts (Campbell and Sedekides, 1999), which increases the challenge of conflict resolution (Paese and Yonker, 2001). Taken together, optimistic overconfidence, confirmatory information search, and the self-threat principle strongly suggest that real players will deviate from Bayesian updating in bargaining and fighting situations, and will do so in an optimistic way.

A well-developed literature, reviewed in chapter 2 and in Samuelson (2001), applies an indirect evolutionary approach to examine what long-run preferences will look like; this approach treats not strategies but preferences as the fixed player traits on which evolutionary selection operates. A consistent finding in this literature is that evolutionary selection for some kind of preferences other than ones for fitness maximization depends on players receiving information about the preferences held by their opponents. The indirect evolutionary approach has been relatively little used in the particular games of political science, and has less often been used to explore the evolution specifically of preferences over beliefs, such as biases towards optimism in the face of bad news. A notable exception is Dickson (2006), who studies a game in which two players must establish mutual trust through costly concessions in order to peacefully divide a pie. In this model, updating beliefs about the opponent's trustworthiness according to Bayes' Rule is shown to yield lower payoffs than adopting a credulous posture in which the opponent is always treated as trustworthy when concessions are made and never when they are not.

3.2.2 Game Theory and Criminal Justice

Several authors have used formal theory to investigate one or more aspects of plea bargaining. Landes (1971) and Adelstein (1978) emphasize that the ability to settle may save both sides in a trial the costs of fighting each other without violating the public's requirements for just sentences on average. Grossman and Katz (1983) and Reinganum (1988) set up games in which prosecutors use plea deals before trial to screen guilty defendants out

from innocent ones when they cannot be perfectly distinguished given the case facts available to the prosecutor. These models have fully separating equilibria in which defendants go to trial if and only if innocent. The later model by Baker and Mezzetti (2001) instead reaches a more intuitive semi-separating equilibrium, where all innocent defendants and some guilty ones reject plea bargains and go to trial; these authors also explicitly model the step at which the prosecutor undertakes investigation before trial, to update beliefs about the defendant's guilt. Gordon and Huber (2002) also includes the investigation step but endogenizes the prosecutor's payoffs in terms of a schedule announced before trial by a representative voter. A commendable review of these and other works on the political economy of prosecution has been undertaken by Gordon and Huber (2009).

One point on which these papers take different approaches is the treatment of the prosecutor's incentives. Most of the game theorists (Grossman and Katz, 1983; Reinganum, 1988; Baker and Mezzetti, 2001) have assumed that the prosecutor has the same utility function as the state and suffers loss from convicting innocent defendants. Other theoretical works (Landes, 1971; Adelstein, 1978; Bar-Gill and Ayal, 2006) have made the contrary assumption that prosecutors seek only to maximize the sentences they secure, regardless of guilt, and this view is backed by several empirical studies (Kessler and Piehl, 1998; Glaeser, Kessler and Piehl, 2000; Boylan, 2005). Gordon and Huber (2002) examine the conditions under which a representative voter can or cannot incentivize a prosecutor to avoid prosecution of the innocent through differential rewards for conviction, dismissal, and acquittal.

3.2.3 Bargaining and Fighting

To date, formal models of plea bargaining have all treated trial as a costly lottery that happens only after the final breakdown of plea negotiations. In practice, bargains may be reached even after the beginning of a trial; opposed advocates may exchange many plea offers over the course of a single trial. The bargaining and fighting that occur are not

one-shot events, but at least finitely repeated.

A class of models using repeated bargaining and fighting has been used in the international relations literature to model interstate war rather than criminal trials. These models emerged as alternatives to the one-shot costly lottery as a description of war. The central insight of these games is that bargaining and fighting complement each other in revealing opponents' private information about their ability to win a whole war. Both offers made and refused at the bargaining table and battles won or lost communicate what opponents know about their own strengths.

Smith (1988) and Slantchev (2003) treat war as a random walk through a unidimensional state space, where total victory is achieved by reaching one end of the space; the latter highlights the opportunity for combatants to screen out weaker opponents earlier in wars using settlement offers acceptable to some but not all types. Filson and Werner (2002) replace the random walk with a limit on the fighting resources of each player, such that each one collapses with certainty after losing some specific number of battles. Powell (2004) considers a sparser formulation in which uncertainty may be over either military strength or the cost of fighting, and simplifies the probability of complete victory in any period to a constant. It is Powell (2004) who makes explicit that players can learn about opponent strength both from bargains offered or rejected and from the chance results of the battles in between the negotiations.

All the bargaining models of war described above take players as perfectly rational in every sense known in the game theory literature. A lone examination of deviation from rationality in one direction is found in Smith and Stam (2004), who relax the assumption of common prior beliefs about while retaining the random walk model of the conflict itself is retained. In contrast to the convergence found in Slantchev (2003), combatants' beliefs can diverge during wartime if they begin with dissimilar priors. Fey and Ramsay (2006) and Smith and Stam (2006) later debate whether or not the relaxation of the common-priors assumption is doing the interesting work in their model.

3.3 The Game

In this game, a prosecutor P and a defense attorney D contest a trial that lasts for two time periods $t \in \{1, 2\}$. In each period, there is a courtroom battle where the defense manages to prevent the prosecution from establishing guilt beyond a reasonable doubt with some probability p (and fails with probability $1 - p$). The periods may be most easily construed as the prosecution's and defense's cases, respectively. If the prosecutor establishes guilt in either period, P wins the trial in that period; the defense is presumed to rest if guilt is established beyond D's power to deny. If the defense prevents establishment of guilt for two periods, D wins the trial. I assume that the prosecutor does not go to trial with a case so weak that the defense can get it dismissed, meaning that the defense cannot win after only one period. The value of winning the trial is taken to be equal to the penalty assigned for the crimes alleged and is normalized to 1.

In each period, prior to the battle in court, the prosecutor makes a plea offer x_t, which the defense may either accept or reject. If the offer is accepted, the trial concludes with a guilty plea to charges incurring a penalty of $1 - x_t$, and the game ends without the occurrence of the period's courtroom battle. If the offer is rejected, the courtroom battle occurs. Every time an offer is rejected, each side pays a cost c, representing the opportunity cost of spending time continuing the trial rather than doing other work. Thus, if an offer is accepted in period t, the payoff to the prosecutor is $U^P = 1 - x_t - (t-1)c$ and the defense's payoff is $U^D = x_t - (t-1)c$; if an offer is rejected in period $t = 2$, the trial winner walks away with $1 - 2c$ and the loser with $-2c$.

Due to disclosure, D knows all the facts the defense and the prosecutor can bring to bear; thus, the defense has perfect information about p. However, even after investigation, if any, the prosecutor does not have perfect knowledge of the defense's ability to win the trial. Defendants who are innocent will have better alibis than ones who are guilty. Although the prosecutor may have learned more about guilt by investigating, ultimately P cannot be as sure whether the defendant is guilty as D can. Therefore, I assume that

the prosecutor knows only that the defense can be one of two types, innocent D^I or guilty D^G,[1] corresponding to different p values $p_I > p_G > c$, and knows the *ex ante* probability q_1 that the defense is innocent.[2]

The prosecutor has a chance to update P's belief about the probability that the defense is innocent between periods of the trial. There are two possible sources of information on which to update. The first is the rejection of the offer itself, since stronger defenses may reject offers that weaker ones would accept. The second is the outcome of the battle. Since the innocent defense has a better chance to survive a courtroom battle than the guilty defense does, the fact of surviving the battle and continuing the game is evidence of a higher probability that the defense is innocent. Since evidence suggesting a lower probability of a innocent defense can only emerge in events that end the game (that is, only when the defense accepts an offer or loses the trial), the prosecutor can only revise q upward during the game. The prosecutor's updated belief at the beginning of the second period will be denoted q_2.

The prosecutor may have either of two learning styles. The "rational" or "smart" prosecutor, P^R, updates beliefs according to Bayes' Rule using all available information. The "optimism-biased" or "dumb" prosecutor, P^B, updates beliefs according to Bayes' Rule using the information inherent in the rejection of an offer, but ignores the fact that a innocent defense is more likely to win a battle. Denote by Ω the rationality level of the prosecutor, so that $\Omega^B = 0$ and $\Omega^R = 1$, and denote by π_I and π_G the probabilities of the innocent and guilty types of defense, respectively, rejecting the offer x_1. Then the prosecutor's posterior belief q_2 is determined as follows:

[1] This is a minor abuse of terminology; the *defendant* is guilty or innocent, but the *defense attorney* is, presumably, not also an accused criminal.

[2] Of course, as Grossman and Katz (1983) state, there are also plausible variations in defense strength that have nothing to do with guilt or innocence. For instance, one defendant may be assigned a lackluster public defender of little legal skill while another is assigned a highly competent attorney. Investigation will not reveal at all whether the accused has a good lawyer. Since the distinction of greatest normative import is that between guilty and innocent defedants, I retain the established description of defense types from the literature.

$$q_2 = \frac{q_1 \pi_I \times (p_I)^{\Omega}}{q_1 \pi_I \times (p_I)^{\Omega} + (1 - q_1)\pi_G \times (p_G)^{\Omega}} \tag{3.1}$$

The previous expression is increasing in Ω. Thus, for a given offer and rejection rate, the smart prosecutor believes more strongly that the defense is innocent than the dumb prosecutor does. This means that, although they start with the same beliefs before the trial, the dumb prosecutor's opinion of the probability of winning the trial stays irrationally high while the smart prosecutor's declines more steeply. Thus, the deviation from rationality exhibited by the dumb prosecutor can be described as optimism bias.

3.3.1 Prosecutors' Incentives and Information in the Model

This game diverges from previous game-theoretic treatments of plea bargaining in two substantive, related ways. First, there is no point in this game where the prosecutor decides whether to investigate, or whether to explicitly drop the case. The most natural interpretation is that investigation, if any, has already concluded and the prosecutor is sufficiently confident of the defendant's guilt to commit to seeking a sentence. Since the prosecutor still faces uncertain information about the defendant, I therefore do not rule out, as previous papers do, the possibility that the prosecutor's investigation provides a false signal of guilt. Baker and Mezzetti (2001) and Gordon and Huber (2002) have modeled investigation, if the prosecutor chooses to undertake it, as always correctly revealing the presence of an innocent defendant, and the latter also supposes that investigation always corectly reveals guilty defendants. This perfect standard seems excessive.

Second, this game does not model the prosecutor as intrinsically concerned with the defendant's guilt or innocence. The prosecutor's payoff is dependent only on the sentence, not the defense's type. Prior models have either endogenized the prosecutor's payoff as a choice made by another player, or assumed that the prosecutor prefers to drop a case rather than convict an innocent defendant. I do not maintain this assumption, for empirical and

theoretical reasons. From the empirical side, Boylan (2005) provides evidence that US Attorneys maximize their chances at desireable future positions, such as judgeships, by maximizing prison months assessed in their offices' cases during their terms, meaning they are incentivized not only to maximally prosecute all defendants regardless of guilt, but also to hire and retain assistants who will do likewise. In the absence of a similiar study of states' chief prosecutors, I assume that a comparable incentive structure applies.

From the theoretical side, there is little reason to think the voting public or other political principals can or will effectively sanction chief prosecutors for trying innocent defendants. Like previous games including Baker and Mezzetti (2001), the (relevant) equilibrium of this game is semi-separating, meaning that all defendants who are innocent reject the initial plea offer, but so do some guilty defendants. Unless all defendants who go to trial are innocent, or the prosecutor's investigations are infallible, the prosecutor always has plausible deniability with regard to having sent innocent persons to trial or prison. Convictions are rewarded, however; Gordon and Huber (2002) show that it is optimal for a representative voter, even one maximally concerned with protecting the innocent, to always vote to re-elect prosecutors who secure convictions. In their model, prosecutors are fully incentivized not to prosecute the innocent only when voters provide maximum feasible rewards for dropped cases, but they show that the voter must be free of any kind of bias against defendants in order to do so.

The bite of these differences from previous models with regard to the evolution of types, however, will be shown to be minimal. For high enough q_1, the behavior and payoffs of the smart and dumb prosecutors will be identical. Thus, the performance of the two types relative to each other would be left unaffected under the contrary assumption that prosecutors drop all cases where the defendant's probability of innocence is higher than some threshold. The features of interest in the model operate under conditions where this model is most like previous ones.

3.4 Solutions to the Game

The appropriate solution concept for this game is perfect Bayesian equilibrium, with the constraint that the innocent type of defense is never more likely to accept a given offer than the guilty type. Of course, the term "Bayesian" may be ill-fitting in the presence of a dumb prosecutor; as described, the dumb prosecutor's belief at time $t = 2$ is not equal to the true posterior probability distribution over types conditional on the history of the game.

I will examine the equilibria under two contrasting assumptions. On the one hand, it may be that the defense knows the prosecutor's learning style with certainty, so that the game is one of only one-sided incomplete information. On the other hand, it may be that the prosecutor's learning style is unknown to the defense, who instead knows only the prior probability ω that the prosecutor is rational. This second case is a game of two-sided incomplete information, in which solutions will be more difficult to find. The most important feature of these solutions will be the difference, if any, between the average payoffs of the two types of prosecutors.

In each case, I assume that the prosecutor knows perfectly whether P is rational or biased, and, importantly, that both sides play the game optimally given everything they know about the prosecutor's learning style. Thus, in period $t = 1$, the biased prosecutor does not act as though P^B will hold the rational prosecutor's beliefs if period $t = 2$ is reached. In this sense, I do not suggest that the biased learning style means knowing in principle what rational updating will consist of but in practice making excuses about why the bad news of losing a battle should be ignored. Rather, the biased prosecutor in the game is fully convinced that P^B knows the right way to learn and that any other style of updating is silly and overly pessimistic.

3.4.1 Screening and Bluffing

At any period, if the prosecutor is sufficiently confident that the defense is innocent, the prosecutor simply offers to give the defense the same amount D^I would win in expectation if the defense refused all offers and fought the trial to the finish. In equilibrium, the defense would accept such an offer whether guilty or innocent. If less sure of the defense's innocence, however, the prosecutor would prefer to make a lower offer; such a lower offer would be accepted only by D^G, who expects lower payoffs from continuing the trial to the bitter end than innocent ones do. Before the last period of the trial, this will let the prosecutor *screen* adversaries to learn more about their guilt, using plea offers.

Since the highest offer any type of prosecutor ever makes is the one that makes the innocent defense exactly indifferent between pleading and fighting on, the strategic calculus of the innocent defense is trivial. However, the guilty defense may have incentive to *bluff* in the first period in a way that convinces the prosecutor to make the maximum offer in the second period. The guilty defense does this with a mixed strategy of accepting the first-period offer just often enough that, if the second period is reached, the prosecutor is indifferent between offering the innocent defense's continuation value and that of the guilty defense. The frequency with which the defense can reject and pull off this bluff will increase as the prosecutor's prior belief that the defense is innocent increases.

If not initially certain enough of the defendant's innocence to prefer avoiding trial altogether, the prosecutor must choose whether or not to make the defense actually do the work of bluffing. The prosecutor can hedge by offering a first-period plea deal equal to the guilty defense's expected payoff assuming D bluffs successfully. This hedging offer is higher than the actual expected payoff if the guilty defense were to contest the whole trial; it reflects the fact that successful bluffing will earn any guilty types who survive one period more than their second-period continuation values. The guilty defense will always accept such an offer; D will have nothing further to gain from actually going through with the bluff.

If even more certain of guilt, however, the prosecutor will lowball the defense and force D to bluff. The lowball offer that would actually make the guilty defense indifferent between bluffing and rejecting outright is lower than the expected payoff to D^G from fighting to the finish. The prosecutor may not get to make an optimally tough lowball offer, however, because the value that makes the guilty defense indifferent between bluffing and rejecting may even be negative; if so, the lowball offer must bottom out at 0, representing no sentence reduction.

Although intuition suggests prosecutors will get tougher on defendants when they are more certain of guilt, the optimal lowball offer gets even lower as the prosecutor's prior belief in the defendant's innocence increases. In a successful bluff, the stronger the defense's position in terms of the prosecutor's prior beliefs, the more often D^G rejects the lowball offer and gets a payoff greater than what D^G would expect from fighting out the whole trial. Thus, as guilt becomes less likely, keeping the defense's payoff to the whole mixed strategy of bluffing equal to that of rejecting outright means making a lowball offer that is increasingly tough. The prosecutor's offer only gets more lenient once P's prior beliefs lean far enough toward innocence that the lowball offer is dominated.

3.4.2 One-Sided Uncertainty

When the defense knows for sure whether the prosecutor is biased or rational, the perfect Bayesian equilibrium is unique for any set of parameters $\{p_I, p_G, c, q_1\}$, and it can be found by backward induction.

Proposition 1. *For any* $\{p_I, p_G, c\}$,

1. *A necessary and sufficient condition for the existence of any* q_1 *where the biased prosecutor's expected utility is less than the rational prosecutor's is* $p_G^2 > cp_G + c$.

2. *A sufficient but not necessary condition for the existence of some* q_1 *where the biased prosecutor's expected utility is greater than the rational prosecutor's is* $p_G^2 \leq cp_G + c$.

Proof. See Appendix 3.A. □

$p_G^2 < cp_G + c$ is simply the condition under which the guilty defense would get a payoff less than 0 from rejecting all offers and contesting the trial to the end. If it is not harder than this to convict guilty defendants, then the prosecutor's lowball offer is bounded below by 0 for any prior beliefs.

How does this work? The guilty defense must work harder to successfully bluff a biased prosecutor than a rational one. The rational prosecutor thus suffers more frequent rejection of the lowball offer than the biased prosecutor does. If the optimal lowball offer is positive, the rational prosecutor can compensate at least somewhat by making a tougher lowball offer. However, if the prosecutor cannot do better than a zero offer, then there is no such compensation.

There are also parameter values for which the guilty defense expects a positive payoff from contesting the whole trial, but the expected payoff is higher for the biased prosecutor than for the rational one:

Proposition 2. *For any* $\{p_I, p_G, c\}$,

- *if there is some* q_1 *for which the prosecutor's best first-period move is the hedging offer, or*
- *if there is some* q_1 *for which the prosecutor's best first-period move is a zero offer,*

then there are some values of q_1 *for which the biased prosecutor's expected payoff is greater than the rational prosecutor's expected payoff.*

Proof. See Appendix 3.B. □

The prior belief threshold at which the prosecutor is indifferent between the lowball offer and the best higher offer is lower for the rational prosecutor than the biased prosecutor, as long as either the best higher offer is the hedging offer or the lowball offer is 0. That is, under either of these circumstances, there must be some prior beliefs for which the biased prosecutor would lowball the defense, but the rational prosecutor would be more

generous. Since both types of prosecutor expect equal payoffs from any first-period move besides the lowball offer, the biased prosecutor must be doing better than the rational one given these priors.

Figure 3.1 illustrates the payoff to each possible offer, for selected values of p_I, p_G, and c, over the range from $q_1 = 0$ to $q_1 = 1$. The biased type's payoffs are indicated with dashed lines; the rational type's payoffs with solid lines. The uppermost of the three lines of each type is that type's overall expected payoff, since the prosecutor chooses an offer that is a best response. Figure 3.2 shows the equilibrium probability of rejection by D^G given the lowball offer against each prosecutor type, and figure 3.3 illustrates the relationship between prior probability of innocence and the optimal lowball offer.

3.4.3 Two-Sided Uncertainty

Now, consider instead the possibility that the defense does not know for certain whether the prosecutor is biased or rational. Instead, the defense knows the prior probability ω that the prosecutor is rational, and can update that belief to ω_1 according to Bayes' Rule after D observes the prosecutor's first-period plea offer. Under these two-sided uncertainty conditions, perfect Bayesian equilibria of the plea bargaining game need not be unique. The decision calculus of the innocent defense is not affected, and the guilty defense's choices will differ only trivially, rejecting the lowball offer as often as against a known rational prosecutor for sufficiently high posterior beliefs ω_1 and as often against a known biased prosecutor otherwise. However, the value of the lowball offer itself need not equal the lowest possible amount that each type of prosecutor could get away with if known. There are many possible combinations of defense beliefs and prosecutor moves that would make each other rational.

To obtain a result regarding the relative performance of biased and rational prosecutors, however, it is not necessary to refine this set of equilibria. Remarkably, in every perfect Bayesian equilibrium of the plea bargaining game under two-sided uncertainty, the

expected payoffs to the biased and rational prosecutors are equal. Of course, if both types of prosecutors prefer to offer first-period plea deals more generous than the lowball offer, their expected outcomes are equal, just as in the case of one-sided uncertainty. What remains to be shown is what happens when at least one of the types prefers to make an offer that the guilty defense sometimes accepts and sometimes rejects.

In any such equilibrium, either both types of prosecutor offer the same first-period plea deal or they at least sometimes offer different ones. If the former, then the initial offer does not allow the defense to update beliefs; the defense chooses a strategy based only on the prior ω. If the latter, then the defense gets to update based on the probability of a given offer being made by a given type of prosecutor before D chooses how frequently to reject the first-period plea deal.

First, consider equilibria with first-period pooling, where both rational and biased prosecutors make the same initial offer. If the defense's prior confidence that the prosecutor is rational is low enough, the defense will accept the first-period plea deal with high enough probability to bluff the biased prosecutor. Since the rational prosecutor believes more strongly than the biased one that the defendant is innocent once the second period is reached, P^R is successfully bluffed whenever the biased prosecutor would be, and makes the same second-period offer as the biased prosecutor. Thus, the two types of prosecutors get the same expected payoff from making a given first-period offer and being treated as the biased type.

On the other hand, if the defense has higher prior confidence in the prosecutor being the rational type, D will reject a common plea deal more often, as appropriate for a known rational prosecutor. In this case, the biased and rational prosecutors will make different offers at time t=2; the biased prosecutor will not be successfully bluffed, but the rational prosecutor will. This means that, in the fraction $q_1 p_I$ of cases where the defense is innocent and survives the first period of the trial, the biased procutor will pay an extra $2c$, the sum of both parties' costs of a second trial period, because the second-period offer will be too

low. In contrast, in the fraction $(1 - q_1)\pi^* p_G$ of cases where the defense is guilty, rejects, and survives the first period, the biased prosecutor will save $p_I - p_G$ in the second period. Substituting in the value of π^* played against a known rational prosecutor yields a net difference of exactly 0 between the two types' expected payoffs. Thus, the biased and rational prosecutors also get the same expected payoff from making a given first-period offer and being treated as the rational type. In this scenario, the biased prosecutor still imagines getting higher payoffs than the rational one would, but the biased prosecutor is incorrect.

Finally, there may be separating or semi-separating equilibria, where the two prosecutor types do not make the same first-period offer. If an equilibrium without first-period pooling by the prosecutor types exists, it still involves the prosecutor types choosing between only two offers: the lowest offer at which D would rather treat the prosecutor as rational rather than always rejecting, and the lowest offer at which D would treat the prosecutor as biased rather than as rational. If such an equilibrium exists, the biased prosecutor must not prefer to be treated as rational any more often than the equilibrium specifies, and the rational prosecutor must not prefer to be treated as biased any more often than the equilibrium specifies. That is, the expected payoff to the biased prosecutor of making the offer associated with being rational must be no greater than the expected payoff from the offer associated with being biased. However, by the preceding results, this means that the expected payoff to the rational prosecutor of being treated as rational must be no greater than the expected payoff from being treated as biased. Since the reverse also holds, the rational and biased prosecutors must get equal expected payoffs even if they do not always make the same first-period offer.

3.5 Career Concerns and Evolution of Types

3.5.1 Evolution of Bias

If cognitive biases such as optimism bias are sticky features of individuals' personality traits rather than strategies they adopt in particular situations, what is the relevance of biased public officials averaging higher payoffs than unbiased ones? Invoking different learning styles' relative performance in the plea bargaining game to predict which types prosecutors are likely to be requires reasons other than prosecutors' strategic choice of their own types.

Chief prosecutors' jobs do not come with tenure. District attorneys in most states are subject to periodic elections; United States Attorneys almost all lose their positions when party control of the White House changes. Many seek other offices, such as judgeships and state-level executive posts, at the end of their terms; others may angle for high-powered or lucrative jobs at prestigious law firms. As noted above, Boylan (2005) finds that the future career attainment of today's chief prosecutors is related to their offices' performance as measured by the total length of sentences the office gets handed down. Chief prosecutors who see current performance as instrumental to their later career ambitions will do all in their power to select and retain assistants those who can optimize that current performance. Assistant prosecutors' prospects for getting hired, and remaining hired, will depend on their ability to extract maximum concessions in bargaining.

Assistant prosecutors, in turn, must at least occasionally decide whether their current jobs are worth keeping to the exclusion of whatever outside options they may have, and they will do so in light of their future career prospects. Assistants may themselves be concerned with moving upto positions as chief prosecutors and beyond. They may also be concerned simply with retaining their own positions when a new boss takes over. All of these goals will be more readily achieved by higher-performing assistants. On the flip side, every year a lawyer spends working in the public sector is a year not spent working

up the ladder in a big law firm. The less past and present performance presages future rewards in the same job, the more tempting an assistant will find it to quit and switch to private-sector work. Thus, assistant prosecutors should themselves choose to retain their positions most highly when they perform well in those jobs.

This intiution, that prosecutors who achieve longer sentences will keep their jobs better, can be translated into a formal model in the terms of the plea bargaining game. Suppose there is a large pool of N prosecutors who, on every date τ, are all busy with cases, and, after each case, let the prosecutor be subject to possibly leaving the job. Let the probability $V(\tau)$ that the prosecutor instead survives to prosecute another case at date be:

$$V(\tau) = \frac{U(\tau) - U^{-}(\tau)}{U^{+}(\tau) - U^{-}(\tau)} = \frac{U(\tau) + 2c}{1 + 2c} \tag{3.2}$$

where $U(\tau)$ is the payoff the prosecutor realizes in the current case, and $U^{-}(\tau) = -2c$ and $U^{+}(\tau) = 1$ are the lowest and highest payoffs, respectively, that any prosecutor could receive in one play of the stage game above[3].

Suppose the prosecutors can be classified into some number of different types. For each type i, let there be $N_i(\tau)$ prosecutors of type i on date τ, where $N(\tau) = \sum_i N_i(\tau)$, and use $V_i(\tau)$ to denote the average survival probability for prosecutors of type i on date τ. $N_i(\tau)(1 - V_i(\tau))$ of type-i prosecutors will leave the pool after each case. By linear approximation, after some small number of cases $d\tau$, the number of type-i prosecutors remaining is approximately

$$N_i(\tau + d\tau) \approx N_i(\tau)(1 - (d\tau - V_i(\tau)d\tau)) \tag{3.3}$$

Use $f_i(\tau) = N_i(\tau)/N(\tau)$ to denote the fraction of the whole pool of prosecutors who are type i and $\overline{V}(\tau)$ to denote the average survival probability among all prosecutors. Then equation 3.3 is enough, using the definition of the derivative as a limit, to derive the

[3]Linearity is not required in the transformation of payoff U to survival probability V; it is only necessary that the transformation by monotonically increasing.

following expression for the time rate of change of $f_i(\tau)$:

$$\frac{\partial f_i}{\partial \tau} = f_i(\tau)(V_i(\tau) - \overline{V}(\tau)) \tag{3.4}$$

This is identical to the *replicator equation*, a canonical and generic game-theoretic model of evolution. Gintis (2009) uses an equivalent derivation to show the replicator equation also obtains if the payoff in the stage game represents the number of offspring an organism is able to produce between one date and the next, or if very long-lived players may randomly learn to play higher-paying strategies with a low probability proportional to the increase in payoff that player would get from switching.

The derivation above suggests that the total number of prosecutors should decrease over time; after the first case there would be $N(\tau)(1 - \overline{V}(\tau))$ fewer than the pool began with. In reality, the total number of prosecutors does not decrease; new recruits fill the jobs of those who leave. If each prosecutor in the pool gets to hire one of the new recruits with equal probability (or if the sub-population of those who are empowered to hire new recruits is representative of the whole pool), and if existing prosecutors select new recruits of the same type as themselves, the replicator equation still obtains with a constant-size pool. These conditions would apply if chief prosecutors who make hiring decisions are more impressed in interviews by candidates who display similar personalities, or if they simply hire their existing friends and tend to make friends with other lawyers who share their personality traits.

The replicator equation predicts that, over time, the fraction of prosecutors who are of some type that averages longer sentences than others should increase, simply because these prosecutors have weaker incentives to quit the job. Thus, under conditions of one-sided uncertainty where prosecutors' learning styles were perfectly known to their adversaries, it would be expected that an initial contingent of biased prosecutors, no matter how small, would inexorably drive rational ones out of the pool. On the other hand, no change in the population fractions of types from initial conditions would be predicted

under two-sided uncertainty.

3.5.2 Evolution of Certainty

Since the payoff advantage to biased prosecutors depends on the non-generic assumption that the defense perfectly knows which prosecutor type is faced, there is no reason to think biased prosecutors will predominate in the long run unless there is also reason to think that one-sided uncertainty actually characterizes at least some prosecutor-defense interactions. In addition to evolutionary foundations of bias, it is necessary to establish an evolutionary foundation for certainty on the part of the defense. One such evolutionary foundation is presented next.

Most evolutionary models make one of two distinct assumptions about the way players are paired against each other in the stage game. One assumption is that of random matching: players have no control over the opponents they face, so every player goes into every game with the same amount and kind of information about their opponent's type. The opposite assumption is assortative matching: players can choose beforehand which type of opponent they are going to face. The situation of a defense attorney taking a case to trial, however, is not well decribed by either of these. On the one hand, the defense obviously cannot choose the prosecutor assigned to the current case. On the other, the defense can at least identify each adversary by name, and the outcomes of that adversary's previous cases will be common knowledge among repeat players in the criminal justice system, if not matters of public record. The defense then can and will begin the game with perfect information about the type of the prosecutor if and only if, in a previous play of the game, the prosecutor has made a move that only P's own type would make.

In two of the possible classes of equilibria of the plea bargaining game under two-sided uncertainty, the rational and biased prosecutors make different moves on at least some occasions, allowing the defense to update beliefs to certainty about the prosecutor's learning style. In equilibria with first-period separation, all biased prosecutors make one offer in

period $t = 1$, and all rational ones make a different initial offer, so the defense always knows the prosecutor's type by the end of the game. In equilibria with first-period pooling and with high enough ω that the guilty defense treats the prosecutor as rational, the biased and rational prosecutors make different offers at period $t = 2$, so that the defense knows the prosecutor's type by the end of the trial in the $q_1 + (1 - q_1)\pi^* p_G$ of trials that reach the second period[4]. For either of these classes of equilibria, then, some prosecutors of publicly unknown type are always being revealed as biased or as rational. Meanwhile, it is also true that new prosecutors are entering the work force as some existing ones leave their jobs. New prosecutors, who have no history of previous cases to their names, must be of publicly unknown type.

The evolutionary dynamic of unknown-type prosecutors revealing revealed as one type or the other and known-type prosecutors being replaced in office by new ones of unknown type will be nested within each of the types themselves subject to the dynamic of the replicator equation. Formally, consider the following generalization of the replicator equation to the case where each type j has some probability ϕ_{ji} of mutating to type i after each play:

$$\frac{\partial f_i}{\partial \tau} = \left(\sum_j f_j(\tau) V_j(\tau) \phi_{ji} \right) - f_i(\tau) \overline{V}(\tau) \tag{3.5}$$

Prosecutors can be grouped into both learning types $\{R, B\}$ and epistemological types $\{K, U\}$ representing whether their learning type is known or unknown. On any date τ, $1 - \overline{V}(\tau)$ are "mutating" from whichever epistemological type they were before to unknown, in the sense that this fraction of each learning type of prosecutors are leaving their jobs and being replaced by new prosecutors with no record. At the same time, for some ϕ determined only by the equilibrium being played, $\overline{V}(\tau)\phi$ of unknown-type prosecutors mutate into known-type prosecutors by making a move that reveals their learning type

[4]The former argument extends to semi-separating equilibria where at least one offer is only ever made by one type of prosecutor; the latter argument extends to semi-separating equilibria where both types at least sometimes make an offer that induces the defense to respond as D would to a known rational prosecutor.

and surviving the current iteration of the game. Given this, and given that all unknown-type prosecutors get the same payoff, the fraction of prosecutors who are biased will still evolve according to the replicator equation, while the fraction of known-type prosecutors evolves over time according to the following equation:

$$\frac{\partial f_K}{\partial \tau} = \overline{V}(\tau)(f_B f_K V_{BK} + (1 - f_B) f_K V_{RK} + (1 - f_K) V_U \phi - f_K) \tag{3.6}$$

The long-run value of f_K, the fraction of prosecutors who are of known learning type, can be found by setting this equation to zero, yielding:

$$f_K = \frac{V_U \phi}{V_U \phi + 1 + f_B(V_{RK} - V_{BK}) - V_{RK}} \tag{3.7}$$

This expression must be greater than 0 for any feasible parameter values other than $\phi = 0$, and must also be less than 1 for any feasible values.

Thus, in the long-run population of prosecutors, there will be some number who have already made clear either that they are rational or that they are biased, as long as an equilibrium is played in which an unknown-type prosecutor has any positive probability of making a move that the opposite learning type would not make. When the defense faces one of these prosecutors, D will be in a position of perfect information about the prosecutor's learning style. In the fraction of trials involving these known-type prosecutors, therefore, the biased ones will still come out ahead, while the expected payoffs in trials involving unknown-type prosecutors will, as shown, be a wash. Since the average payoffs across both known- and unknown-type prosecutors will thus be higher for biased prosecutors, the replicator equation predicts that the proportion of biased prosecutors will monotonically increase over time, at least for a non-trivial subset of the possible equilibria of the plea bargaining game.

3.6 Discussion

3.6.1 Welfare Considerations: Do Biased Prosecutors Hurt the Public?

Would the public benefit if they could curb the evolutionary trend toward optimism bias among prosecutors? That is, are biased prosecutors getting ahead in their careers at the expense of distorting the justice the public would like to impose? In fact, under a set of loose, realistic assumptions about what the public wants out of the justice system, having biased rather than rational prosecutors is in the public interest.

The public gets some positive utility from sentence years assigned to guilty defendants, up to the threshold of the punishment that the public sees as fitting the crime, and some negative utility from sentence years assigned to innocent defendants. The public seeks to minimize the latter, and also seeks to maximize the former, as long as the statutorily assigned penalty for the crime initially charged is the penalty the public actually believes appropriate to that crime. The public may also be concerned with minimizing the amount of time spent in court, since this time may be paid for by taxpayers; however, the nature of the principal-agent relation makes it reasonable to assume that the public places less relative importance on minimizing trial times, and more on optimizing sentences, than the prosecutor does.

First, how do expected sentence lengths and expected trial costs differ between rational and biased prosecutors if the defendant is innocent? In all equilibria of the plea bargaining game, except for equilibria where prosecutors pool on a single first-period offer and the defense treats both types as rational, innocent defendants expect equal-length sentences, and prosecutors expect to spend equal time in court. Outcomes for innocent defendants do differ only if, in the second period, the rational prosecutor makes a plea offer that the innocent defense accepts while the biased one makes a lower offer that is rejected. In that scenario, the actual expected sentence is lower with a biased prosecutor; the defense pays extra trial costs instead of extra sentence years when D rejects the second-period

offer. A public whose utility is decreasing in sentence years for the innocent faster than in the workload of the court would rather innocent defendants see biased prosecutors than rational ones.

Second, how do relevant outcomes differ if the defendant is guilty? Since the prosecutor performs equally well against an innocent defendant regardless of P's type, it must be that biased prosecutors perform better than rational ones conditional on facing guilty defendants if they perform better than rational ones overall. Biased prosecutors, of course, go to trial less often against guilty defendants, meaning they pay lower trial costs and are less often bluffed into an excessively lenient deal with guilty defendants; further, if $p_G^2 \leq cp_G + c$ as in Proposition 1, sentences for guilty defendants are equal conditional on accepting the first-period offer. Thus, under conditions where biased prosecutors would always be expected to out-compete rational ones, guilty defendants average longer sentences and trial costs are lower on average. Both of these are in accordance with the public's assumed preferences.

This section does not exhaust the ways in which the public may be concerned with the rationality of prosecuting attorneys. For one thing, successful prosecutors often continue their careers in still higher offices, including judgeships and state-level executive offices, in which alternate bargaining and fighting need not be a main activity of the job as it is for prosecutors. In other offices to which prosecutors succeed, it is not necessarily true that optimism bias is helpful as it can be in the plea bargaining game. To explore whether or not the public benefits from judges or elected executives who show optimism bias is beyond this chapter's scope, however.

3.6.2 Symmetric Costs and Ultimatum Bargaining

This model highlights a tension between two tenets of conventional wisdom regarding plea bargaining. First, it is thought true that plea bargaining usually results in a reduction of the sentence faced by the defendant, a trade-off by the prosecutor accepted in order

to reduce the work done in prosecuting. Prosecutors who face higher costs for going to trial should thus be willing to make bigger concessions in plea bargaining. Second, it is thought true that plea negotiations mostly involve take-it-or-leave-it offers extended by prosecutors; plea bargaining is ultimatum bargaining, and the prosecutor is always the proposer.

In ultimatum bargaining, however, a successful proposer can always extract a rent from the opponent equal to the entire cost the opponent would pay for failing to reach a deal. The defense cannot get out of paying D's ex ante expected trial costs $c(1+p)$; if D accepts a deal, D simply pays those costs to the prosecutor in the form of additional sentence rather than into thin air in the form of extra work done by going to trial. If the prosecutor's cost per trial period are the same as the defense's, as assumed, then the prosecutor will be fully reimbursed by the defense for the cost of the time P expects to spend in court. The prosecutor does the defense a favor as much as the other way around by reaching a deal, but it is the prosecution who is in position to claim the rewards. Thus, the severity of the plea offer that the prosecutor can get accepted will increase, not decrease, with the cost of time spent in court as compared to the defense's probability of surviving a trial period.

Previous papers in this literature have all modeled the defense as paying no costs for spending time at trial. Doing so is problematic, because it assumes away the principal-agent problem on the side of the defense. The defense attorney, like the prosecutor, has other cases to work on; time spent trying the current one is time not spent acquiring more cases. The defense attorney will face roughly similar incentives to those of the prosecutor with regard to the importance of simply being done with the current case. It will be difficult for the defendant to motivate an attorney to work as hard as possible to minimize the expected sentence under these conditions, even if the defense attorney's compensation is reduced in proportion to the sentence assigned. For faceless public defenders working on salary, even this measure of incentivization is unlikely.

The defendant, however, is the one going to prison; the defendant's cost for spending

time in a trial is presumably negligible compared to what is lost from additional years behind bars. Further, the defendant presumably faces lower opportunity costs; even if acquitted, the defendant probably does not have other trials to go and win. If the defendant is their own agent, rather than strictly a principal represented by an attorney who is an imperfect agent, the prosecutor could not obtain full reimbursement of P's expected trial costs in the plea deal, and the deal might decrease as the trial costs increased relative to the sentence and the probability of conviction.

3.6.3 The Possibility of Citizen Oversight

The principal-agent problem that hinders the state in motivating prosecutors to free innocent defendants is argued above to stem from plausible deniability. The public does not have perfect information on whether a given defendant is innocent any more than a prosecutor does, and will therefore find it hard to sanction prosecutors on the grounds of having tried innocent persons. In at least one situation in the model, however, the innocence of the defendant would be *ex post* certain to a public observer: only innocent defendants ever go the distance at trial. Either of the possible equilibrium second-period offers is generous enough to be accepted by the guilty defendant, but only the higher one also induces the innocent defendant to settle. An interesting extension of the model would impose a small additional penalty on prosecutors who go to a second courtroom battle, representing the small possibility of a public backlash against a prosecutor whom they deduce is trying an innocent person.

The extension with the possibility of a public backlash would have evolutionary implications. The only equilibria in which the lower of the two second-period offers is made are under two-sided uncertainty where the prosecutor is biased, makes the same offer a rational prosecutor would, and is treated as rational by the defense. This would mean that, at least for low enough ω, unknown biased prosecutors would perform worse than unknown rational prosecutors, while known biased ones would still perform better than

known rational ones. Whether the population tended toward more or fewer biased prosecutors would then depend on the magnitude of the possible backlash cost.

3.6.4 The Indictment as Constraint on Offers

It is assumed that offers are bounded by the size of the pie; prosecutors cannot demand that defendants plead guilty to more severe crimes than those on the indictment. There is nothing in the law that requires this, however. Dropping this assumption would mean that, for any parameters, there are at least some values of the prosecutor's prior for which the rational prosecutor outperforms the biased one. The assumption is probably well-founded, however, because plea bargains must be approved by a judge. A bargain that admits more guilt than the indictment alleges runs great risk of being overturned from the bench.

It has also been assumed that the constraint does not bind in the second period. If the second period is reached, no matter what the prosecutor's posterior beliefs, the optimal offer is positive. This assumption is carried over from the analogous models used in international relations, but it need not be appropriate in the criminal context. Convicting a guilty defendant may be so easy that, even after winning one trial period, the guilty defendant's expected payoff for continuing the trial is negative. If this is true, then equilibria where an uncertain defendant treats all prosecutors as rational give the rational prosecutor better payoffs than the biased one. In these equilibria, expected payoffs are equal if each type can make the second-period offer P thinks optimal. However, if guilty defendants are very easy to convict, the biased prosecutor, who is not successfully bluffed, will not be able to make as low a second-period offer as P^B would like; P^B will have to offer 0. This will reduce the amount P^B saves over the rational prosecutor when the defense is actually guilty, without changing what P^B loses compared to the rational prosecutor when the defense is innocent. Dropping the assumption that second-period offers are strictly positive would therefore have evolutionary implications like those of a potential citizen backlash.

3.6.5 Continuous Types

It has been assumed that prosecutors are either fully Bayesian or fully biased. The findings presented here regarding the one-sided uncertainty case should not be altered if, instead, prosecutors may hold any type on a continuum from fully Bayesian to fully biased, as in chapter 2. Where the lowball offer is zero, the proof of proposition 1 serves to show that the prosecutor's expected utility is monotonic in the rationality level Ω, given the parameters of the game; if the lowball offer is zero and P^B performs better than P^R, then P^B also performs better than a prosecutor with any intermediate rationality level $\Omega \in (0,1)$. Likewise, the proof of proposition 2 establishes that the prior probability of innocence q_1 at which the prosecutor is indifferent between x_1^L and x_1^G (or x_1^H, if x_1^G is dominated) is decreasing in Ω, so, for any possible levels of optimism bias, there will be some parameter values for which greater bias provides greater expected utility by inducing a more aggressive first-period offer.

On the other hand, assuming a continnum of prosecutor types somewhat complicates the evolution of certainty for the defense about the prosecution's learning type. Among equilibria in which prosecutors of different types all play the same move in period 1, there are no longer simply some parameter values for which the guilty defense treats the prosecutor as rational (inducing complete separation among prosecutor types in the second period) and others for which the prosecutor is treated as biased; the optimal bluffing probability π^* may be anywhere between the level needed to successfully bluff the rational prosecutor and that needed to successfully bluff the biased one, depending on the game parameters. In these equilibria, no more will be revealed than whether the prosecutor has a rationality level above or below the lowest level for which a prosecutor would be successfully bluffed given the actual value of π^*. For equilibria in which different types make different first-period offers, meanwhile, future defense attorneys will be able to discern the exact type of any prosecutors who previously made the lowball offer, as long as the lowball offer is greater than 0, but will not be able to discern the exact type any prosecutors

who previously offered 0, x_1^G, or x_1^H.

Still, even with only limited information about a given prosecutor's rationality available to future adversaries, the overall evolutionary results will remain. The lower a prosecutor's rationality level Ω is *known* to be, the higher the prosecutor's expected utility; facing a given defendant, a prosecutor for whom Ω is known to be below a given threshold will perform better than one for whom Ω is known to be above the same threshold. There will always be some fraction of prosecutors for whom at least some type information is available, so learning types will do at least some work in determining expected payoffs, and more biased ones will always be favored. Thus, the results presented here that deal directly only with fully rational and fully biased prosecutors are sufficient to summarize what will happen in the less tractable case where the prosecutor may partially learn.

3.7 Conclusion

If they can alternate between offering plea deals and fighting it out in court, prosecutors can benefit from updating their beliefs in a way that makes them excessively optimistic as trials go on. Even if this bias toward optimism is a sticky cognitive trait that prosecutors cannot strategically choose, in the long run most prosecutors will be biased, because their career goals exert evolutionary selection pressure that differentially gets rid of ones who perform below average. The biased prosecutors who emerge, furthermore, prove to be more faithful agents of a public that cares more about assigning just sentences than about reducing the workload of attorneys and the courts.

The game used to model plea negotiations generalizes to other situations of interest in politics. Primary election candidates have opportunities to bargain with each other, trading appointments and other transferable benefits of winning office in exchange for exiting the race, in between episodes where the next state's returns or the next poll updates candidates on how their costly campaigns are performing. In the international relations arena,

alternate bargaining and fighting over a pie characterizes the means by which armed conflicts are resolved; the class of models examined in this chapter was first created to model interstate war. What will differ from context to context, however, is the evolutionary environment in which players find themselves. Factors other than performance in bargaining and fighting games may have much larger selection effects on primary candidates or international leaders than on prosecutors.

3.A Proof of proposition 1

Proof. In period $t = 2$, the defense accepts any offer $x_2 \geq p - c$, and rejects any other offer. Strictly speaking, the defense is indifferent at $x_2 = p - c$, but there exist no equilibria in which D ever rejects this offer.

The prosecutor at time $t = 2$ will offer either $x_2^G = p_G - c$ or $x_2^I = p_I - c$. Any lower offer would result in a jump in expected costs paid, as it would induce one type or the other to reject instead of accepting, and any higher offer would give away unnecessary concessions while saving no expected trial costs. This is the reason why, in equilibrium, it must be that the defense accepts when indifferent; if the defense rejected with any probability when indifferent, no particular x_2 would maximize the prosecutor's utility.

There will be a threshold q_2^* such that, if $q_2 \geq q_2^*$, the prosecutor will offer $x_2^I = p_I - c$, and P will offer $x_2^G = p_G - c$ otherwise:

$$q_2^* = \frac{p_I - p_G}{p_I - p_G + 2c} \tag{3.8}$$

In period $t = 1$, the innocent defense's expected utility for rejecting is dependent only on p_I and c. The innocent defense therefore accepts any first-period offer of at least $x_1^I = p_I^2 - cp_I - c$ and rejects any lesser offer.

The guilty defense, however, may end up playing a mixed strategy. Rather than always accepting or always rejecting x_1, the defense may be best served to accept x_1 just often

enough that, assuming D^G wins the first battle, $q_2 = q_2^*$. That is, the guilty defense may find it useful to bluff in such a way that D^G is treated like D^I in the second period. Of course, if x_1 is high enough that the innocent defense accepts it, the guilty type always does as well, because D^G's continuation value even with bluffing cannot be higher than the innocent type's; thus any offer that is sometimes rejected by the guilty defense is always rejected by an innocent one. By equation 3.8, the probability π^* with which the guilty defense rejects x_1 to set $q_2 = q_2^*$ is therefore:

$$\pi^* = \min\left\{\left(\frac{2cq}{(p_I - p_G)(1-q)}\right)\left(\frac{p_I}{p_G}\right)^{\Omega}, 1\right\} \tag{3.9}$$

π^* is increasing in Ω. Thus, the biased prosecutor is harder to bluff than the rational one; the guilty defense must accept x_1 with higher probability against a biased prosecutor than against a rational one in order to be treated as innocent in the second period. Of course, π^* is also increasing in q_1; the more probably the prosecutor initially thinks the defendant to be innocent, the more often D^G may bluff. $\pi^* = 0$ regardless of Ω for $q_1 = 0$.

If $q_2 \geq q_2^*$, the value to the guilty defense of rejecting x_1 is $x_1^G = p_G p_I - c p_G - c$. For any x_1 less than this, it is preferable to reject with probability π^* rather than always accepting. In the first period, therefore, the prosecutor will choose between offering x_1^I, x_1^G, and a lowball offer x_1^L that the guilty defense at least weakly prefers to reject with probability π^*; the prosecutor cannot gain by making a plea offer so low that both guilty and innocent types always reject it. The minimum such lowball offer the prosecutor can get away with is:

$$x_1^L = \max\left\{\frac{p_G^2 - \pi^* p_I p_G}{1 - q_1} - c p_G - c, 0\right\} \tag{3.10}$$

Since π^* is increasing in Ω, x_1^L is weakly decreasing in Ω. x_1^L is also decreasing in q_1, except where $x_1^L = 0$. If $p_G^2 < c p_G + c$, then $x_1^L = 0$ for all q_1.

The prosecutor's expected payoff from making each of the three possible first-period

offers is therefore:

$$U^P(x_1^I) = 1 - p_I^2 + cp_I + c \tag{3.11}$$

$$U^P(x_1^G) = q_1(1 - p_I^2 + cp_I - c) + (1 - q_1)(1 - p_I p_G + cp_G + c) \tag{3.12}$$

$$U^P(x_1^L) = q_1(1 - p_I^2 + cp_I - c) \tag{3.13}$$

$$+(1 - q_1)(\pi^*(1 - p_I p_G + cp_G - c) + (1 - \pi^*)(1 - x_1^L)) \tag{3.14}$$

The prosecutor's expected utility from the best move in period $t = 1$ is

$$U^P = \max\{U^P(x_1^I), U^P(x_1^G), U^P(x_1^L)\} \tag{3.15}$$

For $q_1 = 0$, $U^P(x_1^L) > U^P(x_1^G) > U^P(x_1^I)$, while for $q_1 = 1$, $U^P(x_1^I) > U^P(x_1^G) = U^P(x_1^L)$. For any p_I, p_G, and c, there exists some $\epsilon > 0$ such that for $q_1 > 1 - \epsilon$, $\pi^* = 1$, and thus, for $1 > q_1 > 1 - \epsilon$, $U^P(x_1^G) > U^P(x_1^L)$. $U^P(x_1^L)$, $U^P(x_1^G)$, and $U^P(x_1^I)$ are all continuous for $q_1 \in (0, 1)$; $U^P(x_1^G)$ and $U^P(x_1^L)$ are decreasing in q_1 for $q_1 \in (0, 1)$ while $U^P(x_1^I)$ is constant in q_1. Thus, for each possible pair of offers, there is exactly one value of q_1 for which the prosecutor is indifferent between the two. For the lowest values of q_1, the prosecutor makes the lowball offer. For the highest values of q_1, the prosecutor offers x_1^I. For some values of p_I, p_G, and c, there is a middle range of q_1 values in which the prosecutor offers x_1^G.

At $q_1 = 0$, both types of prosecutors get the same expected payoff from the lowball offer. If $x_1^L = 0$, the derivative of the lowball offer's expected payoff with respect to q_1 is:

$$\frac{\partial}{\partial q_1} U^P(x_1^L) = -p_I^2 + cp_I - c + \frac{2c}{p_I - p_G}\left(\frac{p_I}{p_G}\right)^{\Omega}(-p_I p_G + cp_G - c) \tag{3.16}$$

This derivative is decreasing in Ω. Thus, for $q > 0$, the biased prosecutor's expected payoff from the zero lowball offer must be higher than the rational prosecutor's. For any

parameter values where the biased prosecutor prefers to offer x_1^L, and where $x_1^L = 0$, the biased prosecutor must get a greater expected payoff than the rational prosecutor.

Finally, if $x_1^L < 0$ at $q_1 = 0$, then $\frac{\partial^2}{\partial q_1 \partial \Omega} U^P(x_1^L) > 0$ at $q_1 = 0$. Thus there exists some $\delta > 0$ such that $U^R(x_1^L) > U^B(x_1^L)$ at $q_1 < \delta$. □

3.B Proof of proposition 2

Proof. If $x_1^L = 0$, the value of q_1 for which $U^P(x_1^G) = U^P(x_1^L)$ is

$$q_1^* = \frac{-p_G p_I + c p_G + c}{-p_G p_I + c p_G + c + \left(\frac{2c}{p_I - p_G} \left(\frac{p_I}{p_G} \right)^{\Omega} (-p_G p_I + c p_G - c) \right)} \tag{3.17}$$

If $x_1^L = 0$, the value of q_1 for which $U^P(x_1^H) = U^P(x_1^L)$ is

$$q_1^{**} = \frac{-p_I^2 + c p_I + c}{-p_I^2 + c p_I - c + \left(\frac{2c}{p_I - p_G} \left(\frac{p_I}{p_G} \right)^{\Omega} (-p_G p_I + c p_G - c) \right)} \tag{3.18}$$

If $x_1^L \neq 0$, the value of q_1 for which $U^P(x_1^G) = U^P(x_1^L)$ is

$$q_1^{***} = \frac{\frac{(p_G^2 - p_G p_I + 2c p_G)}{(2c p_G - 2 p_G p_I + 2c)}}{\frac{(p_G^2 - p_G p_I + 2c p_G)}{(2c p_G - 2 p_G p_I + 2c)} + \frac{2c p_I}{p_G p_I - p_I^2} \left(\frac{p_I}{p_G} \right)^{\Omega}} \tag{3.19}$$

q_1^*, q_1^{**}, and q_1^{***} are all decreasing in Ω. □

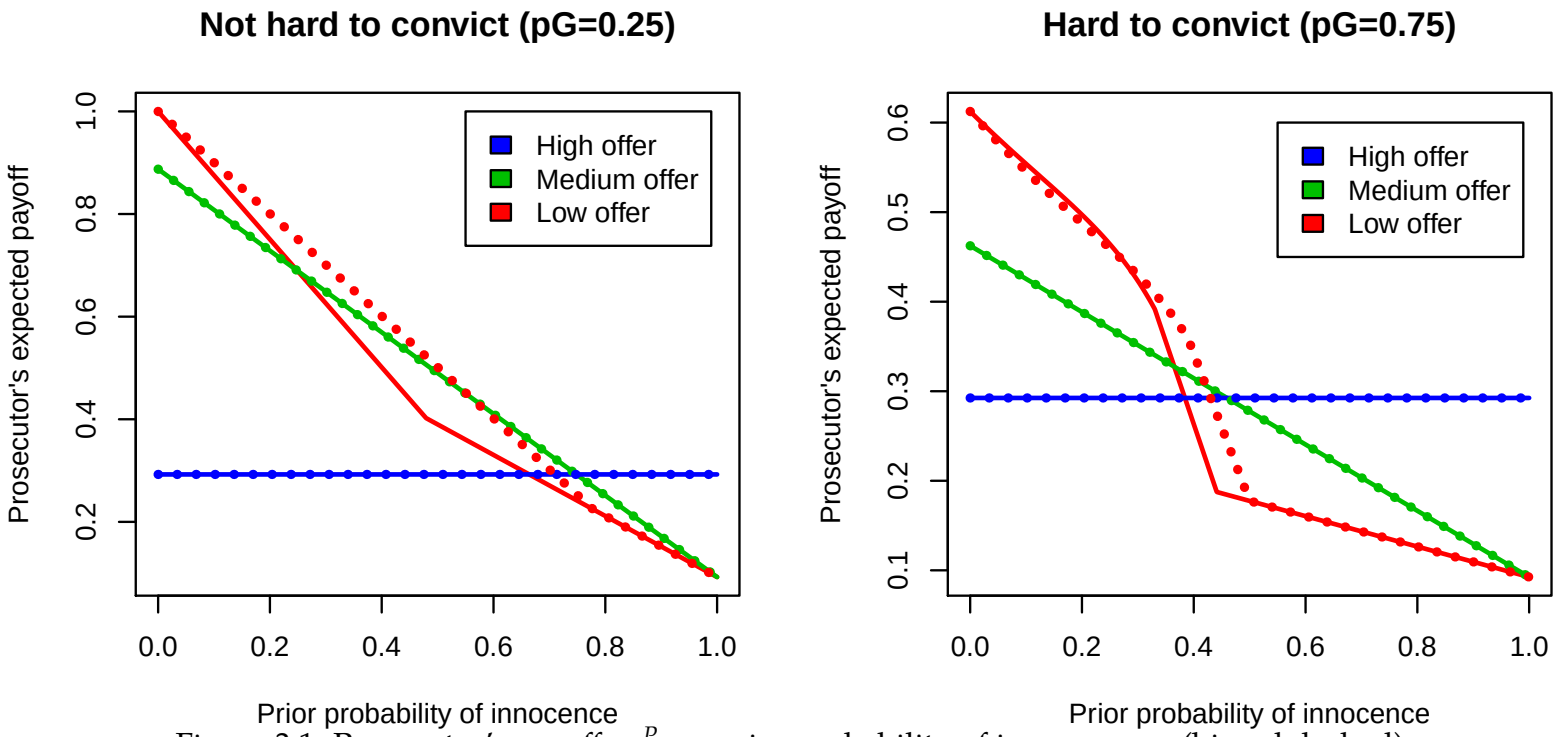

Figure 3.1: Prosecutor's payoffs u^P vs. prior probability of innocence q_1 (biased dashed)

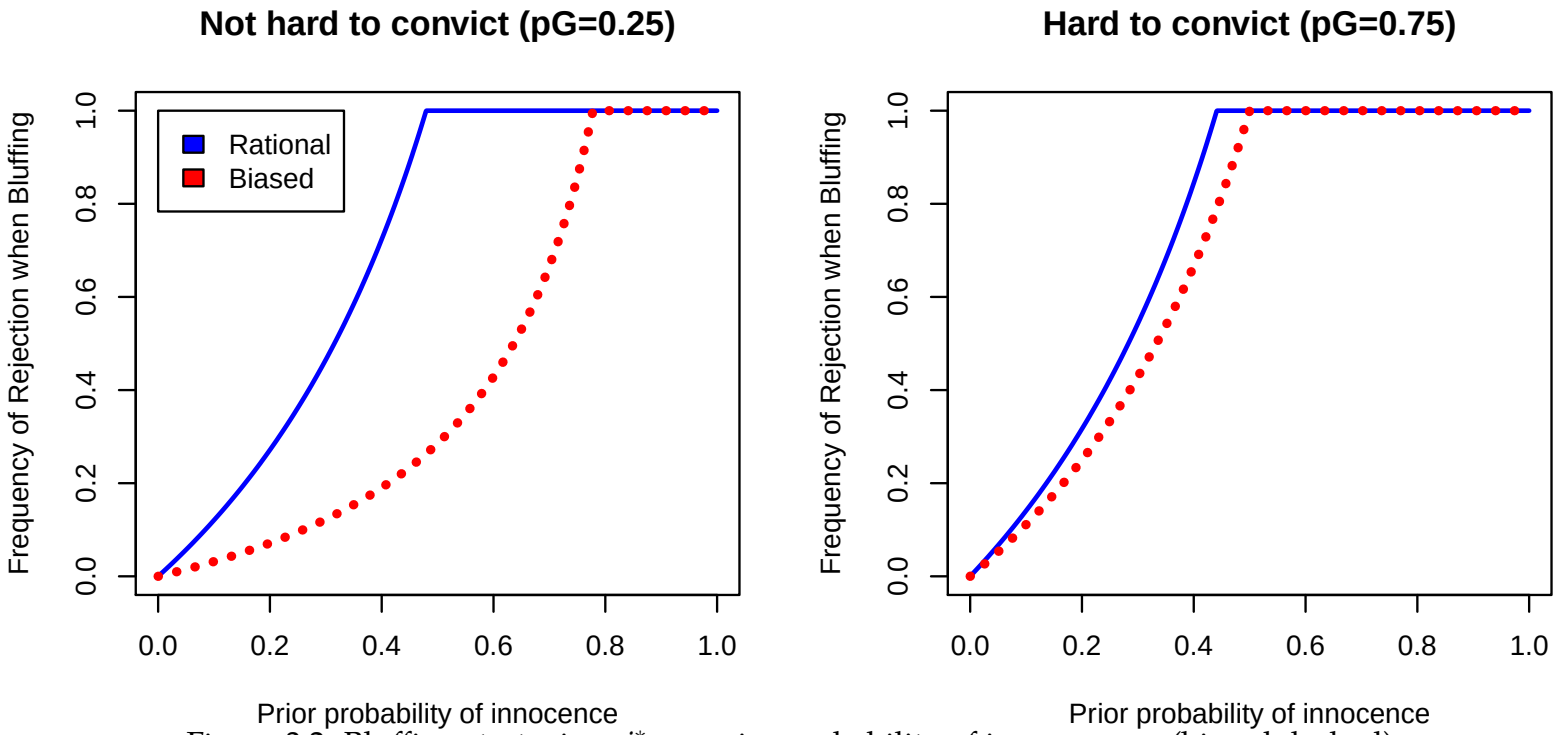

Figure 3.2: Bluffing strategies pi^* vs. prior probability of innocence q_1 (biased dashed)

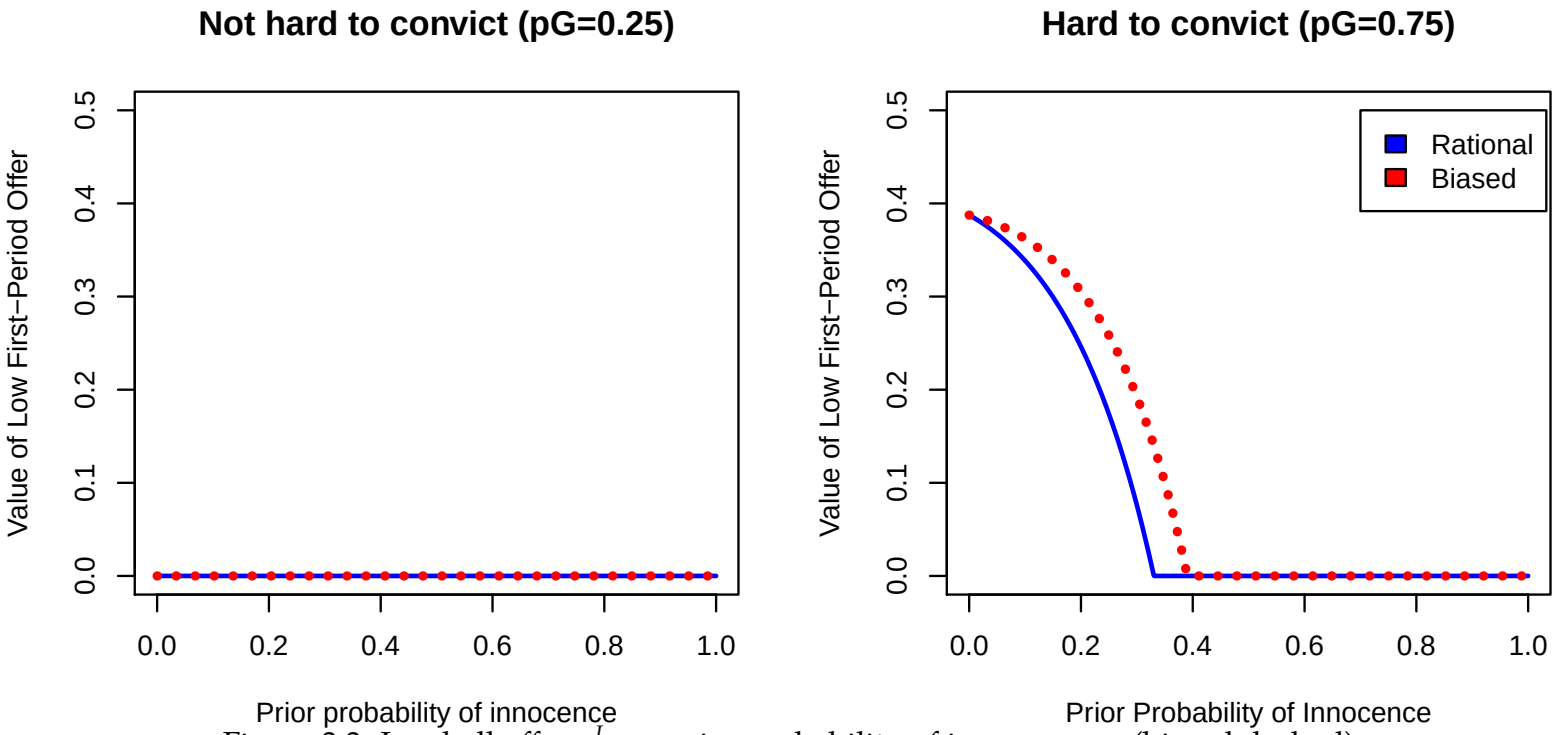

Figure 3.3: Lowball offer x_1^L vs. prior probability of innocence q_1 (biased dashed)

Chapter 4

Ideal Points for All Three Federal Branches Bridged by Interest Group Activity

4.1 Introduction

In order to test theories of how Congress and the President interact with the Supreme Court, it is important that reliable and valid measures of individual ideology comparable across branches are available. Unfortunately for the researcher, since legislators and Presidents do not cast votes on Supreme Court cases and justices do not vote on bills in the legislature, it is not possible to plug data from all three branches into a single model such as NOMINATE. The recent innovation of CFScores (Bonica, 2013) allows ideological scaling of both electoral candidates, including legislators and presidents as well as challengers, and, for the first time, organized interests, using as data on ideological proximity the amount each such interest contributes to each candidate. However, Supreme Court justices do not campaign for election, nor accept contributions from PACs, so they also cannot be dropped directly into CFScores.

In this chapter, I propose to bridge CFScores to vote-based estimates of justices' ideological locations using a source thus far untapped in ideal point estimation: the amicus curiae briefs filed by organized interests who also contribute to campaigns. Hansford (2012) provides the key insight that amicus briefs are measures of ideology for organized interests. Assuming these interests have the same ideological leanings when they act as amici that they have when they act as campaign donors, I treat amici as additional Supreme Court voters in the cases on which they file, and maximize the joint likelihood of votes by justices and amici and of contributions by PACs to candidates over the possible values of the ideal point of each. In this chapter, I present a demonstration of this joint modeling approach using data from the 2006 to 2008 Supreme Court terms and from the 2007-08 election cycle.

4.1.1 What For? Don't We Have a Judicial Common Space?

The current state of the art in locating justices relative to elected political actors is the Judicial Common Space (Epstein et al., 2007). The Judicial Common Space is a mash-up of the dynamic item-response theory model of Martin and Quinn (2002) for judicial ideology with the (first-dimension) Common Space DW-NOMINATE scores of Carroll, Lewis, Lo, McCarty, Poole and Rosenthal (2013) for the ideal points of Presidents and members of Congress. Using fifteen historical instances in which game theory suggests a President had an unconstrained opportunity to place on the Court a new justice located at his own ideal point given the ideal points of Senators, Epstein et al. (2007) estimate a regression of the Common Space score of the appointing President in each instance to the arctangent of the Martin-Quinn score of the justice appointed in the justice's first year on the Court. The arctangent transformation adapts the unconstrained Martin-Quinn space to the $[-1, 1]$ interval on which the Common Space scores live.

While the Judicial Common Space is helpful, it may also be seen to have some shortcomings. First, relative to the size of each of the separate data sets from which the Martin-

Quinn scores and the Common Space scores are estimated, a set of only fifteen bridging observations is fairly small, and invites a search for larger sets of potential bridges. One possible approach to bridging ideal point models with more data, implemented by Bailey (2007), is to consider as votes public statements of position taken by legislators and presidents on Supreme Court cases. However, the whole universe of such position statements will always be more difficult to track down and code than the whole universe of amicus briefs, records of which are maintained by the court itself; the need to search a broad set of sources means that some statements will always be missed, and if there is any relationship between the content of such statements (for example, how extreme the position taken in each statement) and the feasibility of looking them up later, the statement approach may introduce bias.

Second, and more importantly, since the Judicial Common Space is not itself a single model, but a transformation between the point estimates generated by two separate models, it is not possible to completely account for all sources of uncertainty in Judicial Common Space scores. The arctangent transformation uses only the point estimates of both the Martin-Quinn scores and the Common Space scores, meaning that estimates of error in the parameters of the transformation are conditional on the ideal point estimates themselves, and estimates of error in the ideal points are conditional on the transformation and on the other parameters. Additionally, the selection of cases on which to estimate the arctangent transformation is itself conditional on the Common Space scores of Presidents and Senators, taken at their point estimates; the amount of uncertainty about which appointments are unconstrained and therefore define the arctangent transformation is neither well established nor available for incorporation in Judicial Common Space error estimates.[1] By contrast, a true joint model in which ideal points of all parties of interest, together with

[1]It would be possible in theory to generate unconditional standard errors for the whole Judicial Common Space through a clever use of the bootstrap, similar to the use in this chapter, as long as full covariance matrices for both Martin-Quinn and Common Space were available to permit simulation of bootstrap samples of parameters on which to estimate the transformation; however, the error estimates for Common Space scores are themselves from the bootstrap, and neither the bootstrap samples of each Common Space parameter nor a covariance matrix from which to simulate Common Space scores is published.

the parameters that relate them to each other, are estimated at once allows us to say how much we actually know about the relative ideological positions of the politicians under study.

4.2 Model

4.2.1 CFScores

Let $\mathcal{L} = \{1, \ldots, L\}$ be the set of candidates for Federal office and let $\mathcal{K} = \{1, \ldots, K\}$ be the set of all organized interests. Let $y_{tk\ell} \in \{0, \ldots, 10\}$ be the contribution, denominated in units of \$500, given by each interest $k \in \mathcal{K}$ to each candidate $\ell \in \mathcal{L}$ during each phase $t \in \{p, g\}$ of the election cycle (the primary and general elections, respectively). Let each candidate ℓ have an ideal point θ_ℓ and let each interest k have ideal point δ_k. Let $\boldsymbol{X_{tk\ell}}$ be a set of covariates describing non-spatial characteristics of each candidate ℓ and their relationship to each interest k in each phase t of the election cycle, and let $\boldsymbol{\zeta_k}$ be a vector of coefficients describing the effect of each covariate on giving by each group k. Finally, let α_k and γ_ℓ be interest- and candidate-specific intercepts describing overall propensity to give and to receive campaign contributions, and let $\sigma_k > 0$ be a shape parameter representing the similarity of the dispersion of interest k's gifts to that of a Poisson distribution.

Using these parameters, Bonica (2013) gives the following likelihood for $y_{tk\ell}$:

$$h(y_{tk\ell}|\boldsymbol{\delta, \theta, \alpha, \gamma, \sigma, \zeta}, \varphi; \mathbf{X}_{tk\ell}) = \begin{cases} \text{NB}\left(y_{tk\ell} \left| e^{\left(\alpha_k + \gamma_\ell + \varphi(\delta_k - \theta_\ell)^2 + X'_{tk\ell}\zeta_k\right)}, \sigma\right.\right), & y_{k\ell} < 10 \\ 1 - \sum_{y'=0}^{9} h(y'|\boldsymbol{\delta, \theta, \alpha, \gamma, \sigma, \zeta}, \varphi; \mathbf{X}_{tk\ell}), & y_{k\ell} = 10 \end{cases} \tag{4.1}$$

where $\text{NB}(y|\lambda, \sigma)$ represents the probability mass function at y of the negative binomial distribution with mean λ and shape parameter σ.[2]

CFScores is a count regression, using a negative binomial link, for the amount given

[2] Bonica (2013) sets $\varphi = 1$ and fixes one ideal point to identify the scale of the ideal points; I fix two interest group ideal points and let φ vary. Either definition is sensible.

by an organized interest group to a candidate for office in an election cycle as a function of the distance between them in an ideological space and of other, non-spatial candidate characteristics, such as leadership, seniority, geographic colocation, common legislative interests, and the candidate's prospects of election. The importance of each non-spatial candidate trait is estimated separately for each interest group, meaning that the number of parameters to be estimated is very large; however, the model thus accounts for the ways in which campaign contribution strategy varies across organized interests. The model accounts for limits imposed by campaign finance law by treating contributions of \$5,000 as right-censored, and accounts for overdispersion in gifts at the level of each organized interest as well.

The negative binomial count model is derived from a theoretical picture of organized interests as maximizing the utility they obtain from making contributions, across all possible recipients, given a finite budget for campaign spending. If the budget constraint is assumed to bind exactly, contributions are not strictly count data but compositional data, requiring in principle a more complex model accounting for negative correlation between gifts to each candidate and gifts to each other; the model more accurately depicts, however, the situation in which interests operate far from their absolute spending limits (which are unknown) but instead make contributions until the marginal utility of giving another dollar is equal to that of another dollar left in the organization's coffers. McCarthy, Chen and Smyth (2012) apply an analogous negative binomial model to count-like compositional data from RNA sequencing to estimate how changing conditions affect expression of each of an organism's genes, although differential gene expression does not involve censoring or an ideological similarity predictor.

4.2.2 Justice votes and amicus briefs

Let $\mathcal{J} = \{1, \ldots, J\}$ represent the set of justices and $\mathcal{I} = \{1, \ldots, I\}$ the set of cases. For each $i \in \mathcal{I}$, let $\mathcal{K}_i \subseteq \mathcal{K}$ be the subset of interests who also served as amici on case i. Let

$z_{ij} \in \{p, r\}$ represent the vote each justice $j \in \mathcal{J}$ casts in each case $i \in \mathcal{I}$, where p is a vote that favors the petitioner and r favors the respondent. Similarly, let $w_{ik} \in \{p, r\}$ represent the amicus brief filed by each interest $k \in \mathcal{K}_i$ in each case $i \in \mathcal{I}$, where p and r represent briefs supporting the petitioner and respondent, respectively. Let ω_j be the ideal point of each justice j, and let η_i^p and η_i^r be the ideal points of the petitioner and respondent, respectively, in each case i. Finally, let $\xi > 0$ and $\beta > 0$ be amplitude parameters describing the extent to which ideology determines vote choice for amici and justices, respectively, and let $\tau > 0$ be a width parameter describing the shape of the function mapping the spatial distance between a voter and a case party to the utility of voting for that party.

Using the above parameters, I assume the following likelihood functions for w_{ik} and z_{ij}:

$$f(w_{ik}|\boldsymbol{\delta}, \boldsymbol{\eta}, \xi, \tau) = \begin{cases} \Phi[\xi(\phi[\delta_k - \eta_i^r, \tau] - \phi[\delta_k - \eta_i^p, \tau]), 1], & w_{ik} = r \\ \Phi[\xi(\phi[\delta_k - \eta_i^p, \tau] - \phi[\delta_k - \eta_i^r, \tau]), 1], & w_{ik} = p \end{cases} \tag{4.2}$$

$$g(z_{ij}|\boldsymbol{\omega}, \boldsymbol{\eta}, \beta, \tau) = \begin{cases} \Phi[\beta(\phi[\omega_j - \eta_i^r, \tau] - \phi[\omega_j - \eta_i^p, \tau]), 1], & z_{ij} = r \\ \Phi[\beta(\phi[\omega_j - \eta_i^p, \tau] - \phi[\omega_j - \eta_i^r, \tau]), 1], & z_{ij} = p \end{cases} \tag{4.3}$$

where Φ is the normal (Gaussian) cumulative distribution function and ϕ is the PDF of the same distribution.

This is a binary voting model like that of Martin and Quinn (2002), in which any amici who participate in a given case are assumed to be additional voters, but in which each voter's expected utility of voting for each case party is assumed to follow a Gaussian rather than quadratic function of the distance between the two. Below, I elaborate on the choice of a Gaussian utility function and the absence in the model of an abstention option for potential amici.

I do not model justices' ideal points as changing over time because, as discussed in section 4.3, the data cover only a short period of time. If estimates were to be produced

for several years, it would make sense to incorporate a random-walk prior on the justice ideal points as in Martin and Quinn (2002).

4.2.2.1 Amicus vote choice and participation

As noted above, organized interests must decide not only which side to support in a Supreme Court case but whether to participate as amici at all. Filing a brief is costly, so most interests abstain from most cases. It may thus appear desirable to model both participation and vote choice when modeling amicus activity. In addition to the model reported, I also fit an ordered model treating non-filing as a middle category between supporting the petitioner and supporting the respondent, becoming more likely the smaller the absolute difference in an interest's spatial utility from supporting each party. I assumed each amicus in the data set had abstained from filing in every case on the same issue, as coded by Spaeth, as a case in which they did file a brief, giving a data set with 317 examples of meaningful abstention from filing in addition to 201 briefs filed. The ordered model yielded justice ideal point estimates that were substantively no different from those estimated using a vote choice model only, but was estimated with substantially greater difficulty. Because the binary and ordered models yielded such similar estimated of quantities of interest, and because predicting amicus participation is not itself a principal goal of this project, I believe the model of vote choice only is a better option for bridging justice and candidate ideology models.

4.2.2.2 Gaussian utility

I follow Poole and Rosenthal (1985) and assume that the spatial component of the utility a given justice or amicus gains from voting for a given case party is Gaussian in the distance between their ideal points.[3] This is in agreement with the findings of Carroll, Lewis, Lo,

[3] An analogous model assuming quadratic utility has also been estimated; it neither broadly contradicted the results presented below nor alleviated the computational difficulties associated with getting error measures. Surprisingly, the model is also at least partly robust to the unusual assumption of hyper-Gaussian

Poole and Rosenthal (2013) for a variety of voting bodies including the Supreme Court, and is theoretically appealing. Under Gaussian utility, a small difference between the distance from a voter's ideal point to one option and the distance to the other option has a smaller impact on vote choice when both options are far from the voter's ideal point. As both case parties move sufficiently far from the ideal point of a given justice, spatial proximity to each party will give way to factors not modeled here in the justice's (or the amicus') decision calculus. Thus, a case such as *Bush v. Gore* might be expected to elicit more ideologically driven votes than a hypothetical case on the same issue with either less extreme case parties (say, *Dole v. Clinton*) or more extreme ones (*Franco v. Tito*).

Additionally, Gaussian utility is required for model variants in which it is assumed that organized interests may abstain from filing amicus briefs, as described above. With quadratic utility as assumed in Martin and Quinn (2002), it would be possible to assume that abstention is likely to occur when both case parties occupy similar ideological locations or that abstention is likely when both case parties are extreme, but not both. The former could be accomplished by treating abstention as a response ordered strictly between supporting the petitioner and supporting the respondent; the latter would come from a multinomial specification with abstaining assumed to provide a constant utility not dependent on the case parties' locations. By contrast, the ordered model can accomodate abstentions for both reasons if utility is assumed Gaussian in ideological distance. Since, as noted, it appears sensible not to model interests' decision between participation and abstention after all, it may be worth in depth comparing models with Gaussian and quadratic utility specifications for justices and amici in future iterations of this work, but the Gaussian specification will suffice for now.

I assume a common width parameter for the Gaussian utility functions of amici and justices. This assumption is consistent with the finding of Hansford (2012) that justices are more likely to vote ideologically in cases where more amicus briefs are filed; there would

spatial utility, proportional to $e^{-((x^2)^2)}$ for ideological distance x, rather than Gaussian. Model versions with hyper-Gaussian utility and an abstention option do not yield sensible results, however.

be less reason to expect this finding if justices and amici had utility functions of substantially different shapes. However, I allow the amplitude of the Gaussian utility function to take on different values for justices and for amici. Justices are likely to be far more concerned with non-spatial issues of law than those who select into amicus participation are, so the spatial component of utility should be a weaker predictor of vote choice for justices than for amici.

Meanwhile, Gaussian utility is not assumed to apply in the model for campaign contributions; the log expected donation is assumed to fall off quadratically in the distance between a given interest and candidate. This is a matter both of computational convenience – assuming a likelihood function not everywhere concave in all its parameters for the campaign finance data set would frustrate optimization much more than doing so for the relatively small set of data on Supreme Court outcomes – and also of theoretical appropriateness. Campaign contributors are choosing not which of two alternatives, perhaps both so ideologically distant as to be indistinguishably morally odious, to support, but how much support to give to a single recipient, where even the amount of support given to competing recipients is essentially independent. The prospect of indifference between choices via alienation from all of them, which should lead to votes predicted worse by ideological distance as distance increases, does not apply to the campaign contribution scenario.

4.3 Data

Data on campaign finance are the replication data from Bonica (2013); data on justice votes and on Supreme Court cases are from the Supreme Court Database; data on amicus activity come from those used in Hansford (2012). To prevent data size from slowing model development, I use data from a single time period: the 2007-08 election cycle, during which the 110th Congress sat, and the 2006 through 2008 Supreme Court terms.

4.3.1 Campaign finance data

The campaign finance data record the contributions (if any) of each of 1,067 organized interests to each of 654 candidates for Federal office in the primary and general election phases of the election cycle. The candidates include all registered candidates who received at least 25 nonzero contributions from included contributors during the election cycle: 541 House candidates, 106 Senate candidates, and seven candidates for the Presidency. The contributors are all organized interests who gave at least 25 nonzero contributions to included candidates.

In addition to the amount, coded in $500 units, of each contribution, each data point includes 16 dummy variables used to control for non-spatial candidate characteristics that may drive contributions from organized interests. Cases where one of these covariates perfectly predicts zero contribution by a given interest are dropped, as are primary election data for candidates who received gifts only in the general election phase or vice versa, leaving 1,248,322 total data points.

I inherit directly from Bonica (2013) controls for: candidacy for a seat in each interest's home state; Senate candidacy; Presidential candidacy; incumbency; candidacy for an open seat; freshman status among incumbents; membership on each of four key committees (Appropriations, Ways and Means, Energy and Commerce, and Banking); leadership of a party or a committee delegation[4]; and the phase of the election cycle.

The remaining covariates are modified slightly from Bonica (2013). To conserve a few degrees of freedom, I collapse the eight margin-of-victory bins that Bonica (2013) uses to control for electoral competitiveness into four, creating dummy variables for winning more than 55% of the vote, between 45% to 55%, and between 30% and 44%, with a residual category for winning fewer votes or failing to run despite receiving contributions. I also include a dummy for membership on a committee relevant to the work of each organized

[4]The text of Bonica (2013) describes these as two separate variables, but the replication code creates a single dummy for both, and results are reported for a single dummy.

interest, which is coded in the replication data provided with Bonica (2013) but is omitted from the text.[5]

Coefficients were not estimated for covariates that perfectly predicted either a zero gift or a maximum gift by a given interest group. Data points for which the candidate's value on one or more covariates perfectly predicted a zero or maximum gift were dropped from the set, equivalent to treating these perfectly predicted contribution amounts as probability-1 events.

4.3.2 Judicial data

The three Court terms covered fall within a single natural court under Chief Justice John Roberts; the nine justices cast 1464 votes on 164 cases in this data set, with 12 instances where a justice did not cast a vote due to recusal or for other reasons. In addition to these 1464 votes, the data used include 201 amicus curiae briefs filed in support of one side or the other in the same case. This includes every brief filed in support of either the petitioner or the respondent[6] by each organized interest that also appears in the campaign finance data used, except for corporations; the latter are dropped as likely to have substantial non-ideological, business-specific motives for filing. 16 of the cases in the data set were decided unanimously; in these cases, the only support for the losing side was from one or more amici curiae.

4.4 Numerical Methods

I coded the model's log likelihood, and a finite-difference approximation to the gradient of the log likelihood, in `C`, and then, starting from the initial values described in 4.4.1,

[5] Bonica (2013) also controls for budget size as measured by the total amount of money given to all candidates by all interests in a given election cycle; this control is not meaningful in the current data set, since only one election cycle is under consideration.

[6] A small number briefs are coded by Hansford (2012) neither as clearly supporting the petitioner nor as clearly supporting the respondent, and are omitted.

maximized the log likelihood in `R` using the `L-BFGS-B` method in the `optim()` function. As discussed in 4.4.2, I then generated 95% confidence intervals from the parameteric bootstrap.[7]

4.4.1 Initializing Parameter Values

The model was optimized in two stages. In the first, I initialized the candidate and contributor ideal points using the iterated money-weighted averaging procedure prescribed in Bonica (2013), and initialized the remaining CFScores parameters using negative binomial regressions for each contributor without accounting for censoring; leaving the ideal points for Americans for Democratic Action and the National Conservative Campaign Fund fixed at their initial values (of -1.532 and 1.116, respectively) to anchor the left and right ends of the scale, I then estimated CFScores only starting from those initial values. In the second, I initialized the ideal points of Justices Breyer, Ginsburg, Souter, and Stevens to the estimated CFScore of the mean Democrat in the data set and those of the other five justices to that of the mean Republican, initialized case party locations to the mean location of justices and amici supporting them, and optimized the likelihood of the full model starting from those locations and the CFScores-only estimates of the other parameters.[8]

4.4.2 Confidence Intervals From the Parametric Bootstrap

Because of the Gaussian rather than quadratic utility function, the likelihood is not guaranteed to be globally concave; for that reason, and because of the limited precision to which

[7] As with most output in the ideal point estimation literature, I do not deal directly with the potential for inconsistency noted by Londregan (1999) when the number of parameters grows with the number of data points. This inconsistency arises because, given finitely many voters with fixed ideal points, only a finite set of values could be the maximum likelihood estimate of the location of an alternative and vice versa. However, with large numbers of alternatives and voters each, the granularity imposed by the finiteness of each set will be small, so researchers estimating ideal points do not make an issue of it. I thank a commenter at the 2014 Annual Meeting of the Midwest Political Science Association for pointing out the existence of this inconsistency result.

[8] In the tradition of "Poole and Rosenthal as the outer loop of the estimation" (Carroll et al., 2009), the initial values thus chosen are considered acceptable on the grounds that there are no obvious anomalies in the final results; robustness to other initial values is not guaranteed.

I ran the numerical optimizer and the large number of parameters, a positive-definite approximate Hessian for the whole model's negative log likelihood is not available.

Instead of an inverse Hessian, I use the parametric bootstrap (Efron and Tibshirani, 1994) to generate confidence intervals for quantities of interest from the model. The parametric bootstrap is a resampling procedure in which the original predictor variables and the maximum-likelihood estimates of the parameters are used to simulate new sets of outcomes from the model, and the model is then re-estimated on each new set of outcomes to obtain a sample from the approximate distribution of the parameters. I conducted 1,057 bootstrap samples of the parameters[9].

The parametric bootstrap is appropriate for cases in which the data represent the universe of possible cases (a data point for every candidate-contributor dyad, for every justice-case dyad, etc.) rather than a random sample of the possible cases (which would lend itself to the nonparametric bootstrap). The parametric bootstrap is also useful specifically where the likelihood is not concave everywhere, since a numerical optimizer need not converge to the same local maximum of the likelihood function on every randomly simulated data set that it converged to on the original data. Bootstrap methods can make efficient use of parallel computational resources with minimal added programming effort. Lastly, the bootstrap makes it easy to calculate confidence intervals for quantities not estimated in the model itself, such as the median ideal point within a chamber or party. For reasons like these, the parametric bootstrap is already applied by Lewis and Poole (2004) to generate the standard errors for NOMINATE.

The bootstrap is not without its caveats. The amplitude parameters φ, β, and ξ measure, in part, how successfully ideological distance (relative to the distance between the fixed interest groups) predicts contribution and voting outcomes. Since the model (in expectation) fits data simulated directly from the model better than it fits real-world data (on which it suffers at least some omitted-variable bias), bootstrap values of parameters

[9]The intended total was 1,125 – 125 on each of six processor cores of one computer and each of three of another – but a power outage interrupted computation on the latter machine shortly before completion.

that correlate with the model's predicive ability will be higher on average than the maximum likelihood estimate of these parameters obtained with real data. In other words, the parameteric bootstrap introduces a form of bias owing to the parameterization of the model. Confidence intervals for these less substantive parameters are not computed; however, bias is a concern for parameters of interest as well, not least because they depend on the amplitude parameters.

To deal with possible bias, I incorporate the bias correction specified by Efron (1987) in defining the BC_a bootstrap confidence interval. However, rather than estimate the computationally expensive acceleration parameter, I assume it is zero. The acceleration adjusts for the dependence of the standard error of a parameter θ on the value of θ; for the quantities of interest, $\text{se}(\theta)$ and θ should be independent as a matter of theory. The acceleration is calculated using the jackknife, the re-estimation of the model once for each data point on the entire set of data less that point; with over 1.2 million rows in the campaign finance data set alone, applying the jackknife to this problem is infeasible.

Other approaches to confidence interval generation in the absence of an invertible, positive-definite numerical Hessian are available, but present their own challenges. Importance sampling as suggested by Gill and King (2004) is infeasible due to the large number of parameters in the model. Markov chain Monte Carlo approaches to sampling are more promising in theory, relying neither on a numerical gradient or Hessian nor simulated data, but would require a large amount of time to explore the whole parameter space. Furthermore, since exact conditional distributions of each parameter are not available, the MCMC algorithm used would have to include a Metropolis step, in which proposed new values for one or more parameters at a time are drawn from some distribution, and the Markov chain moves to the proposed ones with a probability dependent on the posterior at the proposed and current parameters. However, choosing a suitable proposal distribution is itself difficult when the target distribution is not reasonably approximated by a normal distribution, as is true of the distributions of justice and case party ideal points.

Lastly, diagnosis of convergence for a model containing as many parameters as are present here will be laborious. For now the parametric bootstrap remains the best way to generate confidence intervals for this problem.

4.5 Results

The results overall depict the justices of the 2006-08 period as quite moderate compared to politicians campaigning for Federal office at the same time. Table 4.1 and figure 4.1 show estimated ideal points, and bias-corrected 95% bootstrap confidence intervals for these ideal points, for each justice, for the median of each of eight subsets of candidates for Congressional seats[10], and for each of the seven Presidential candidates for whom positions could be estimated. While the Court's liberal and conservative wings are visibly, significantly separate from each other, all justices' ideal points are located left of the 95% confidence interval for the median Republican candidate, and all justices' ideal points (except, just barely, John Paul Stevens) are located right of the 95% confidence interval for the median Democratic candidate.

While the model supports the conclusion that justices are centrist, it does not support concluding that they are not ideological; ideological distance correctly predicts 89.7% of justice votes and 80.8% of amicus votes. Rather, justices appear to experience very strong motivations to vote for case parties located within a small range centered on their own ideal points and to be essentially indifferent to ideology for case parties outside that range. The standard deviation τ of the Gaussian utility curve for justices and amici is estimated at 0.233, while the utility amplitude ξ for amici is estimated at 10.900 and the amplitude β for justices at a staggering 48.925.[11] That the extent to which justices appear motivated

[10] I do not estimate the more intuitively interesting median *incumbent* in each category, because some incumbents in the 110th Congress did not receive enough contributions during the 2007-08 election cycle to estimate ideal points for them.

[11] A confidence interval, even bias-corrected, is not available for β, as the maximum likelihood estimate on the real data was lower than all of the bootstrap estimates.

Name	Ideal Point	(2.5%,	97.5%)
Justices			
Stevens, John Paul	−1.166	(−1.470,	−0.938)
Ginsburg, Ruth Bader	−1.059	(−1.279,	−0.820)
Souter, David	−1.024	(−1.318,	−0.823)
Breyer, Stephen	−0.832	(−1.043,	−0.622)
Kennedy, Anthony	−0.197	(−0.340,	0.022)
Roberts, John	−0.100	(−0.247,	0.127)
Alito, Samuel	−0.081	(−0.221,	0.171)
Scalia, Antonin	−0.030	(−0.173,	0.208)
Thomas, Clarence	0.115	(−0.033,	0.445)
Median Candidates			
Democratic House	−1.315	(−1.488,	−1.088)
Democratic	−1.306	(−1.465,	−1.072)
Democratic Senate	−1.116	(−1.184,	−0.956)
Senate	−0.786	(−0.950,	−0.643)
House	−0.399	(−0.512,	−0.333)
Republican House	0.271	(0.105,	0.391)
Republican	0.312	(0.173,	0.411)
Republican Senate	0.472	(0.351,	0.551)
Presidential Candidates			
Clinton, Hillary Rodham	−2.756	(−3.259,	−2.458)
Richardson, William B.	−1.620	(−1.842,	−1.210)
Dodd, Christopher J.	−1.521	(−1.753,	−1.343)
Romney, Mitt	−0.577	(−0.896,	0.022)
Bayh, Evan	−0.239	(−0.551,	0.194)
Giuliani, Rudolph W.	0.606	(0.346,	1.011)
McCain, John S.	2.050	(1.639,	2.577)

Table 4.1: Locations of justices, median legislative candidates, and presidential candidates.

by ideology is greater than that for those who have selected into amicus participation is surprising; that justices more predictably vote for ideologically nearby case parties than amici curiae do calls into question the extent to which justices are constrained by law in every case.

Figure 4.2 illustrates the stark relationship between justices' votes and ideological location. I plot the probability of a liberal vote as a function of justice ideological location, on a hypothetical case where the parties are separated by the median petitioner-respondent ideological distance and centered at the median case party location; hypothetical case

party locations are shown by dashed lines and the estimated location of each actual justice is shown by a faint dotted line. The transition from the range of ideal points in which a justice would be essentially guaranteed to vote for the liberal outcome in this hypothetical case to that in which a conservative vote would be essentially guaranteed is extremely narrow, while each such range is approximately as broad as the range between the most conservative and most liberal justice; all justices fall squarely within one of the regions in which probability of voting for the closer party is indistinguishable from 1.

An in-depth comparison between the scores estimated here and existing sets of ideology scores, such as the Judicial Common Space, would be premature, since such a restricted set of data has been used here. Nonetheless, as a sanity check and a demonstration of techniques for comparison of this model as estimated on a complete data set to others, table 4.2 shows where each justice ranks in conservatism among the 521 incumbent legislators who were scored in both this model and the Judicial Common Space (using JCS scores for 2007), with bias-corrected bootstrap confidence intervals for the former. The models capture approximately the same ordering of justices; the JCS reverses this model's ordering of Justices Roberts and Alito, but those two are statistically indistinguishable from each other, and even Justice Breyer, placed left of Ginsburg and Souter by the JCS and right of them by this model, has a 95% confidence interval that includes both of their ideal points. The Judicial Common Space, meanwhile, supposes the justices as a group to be more extreme compared to other politicians than this model does. The null hypothesis that the two models locate each justice to the left of equally many legislators is rejected with 95% confidence for seven of nine (all except Roberts and Alito), and for six of those (the exception being Kennedy), the Judicial Common Space provides the more extreme estimate of their ideology. Compared to this model, the JCS places the members of the court's liberal wing closer together and the members of its conservative wing much further apart. While this model's emphasis on the overall centrism of the justices compared to the Judicial Common Space might be read to suggest only strong dependence on the

Justice	Rank	(2.5%,	97.5%)	JCS
Stevens	382	(334,	458)	479
Ginsburg	351	(309,	412)	445
Souter	342	(310,	430)	442
Breyer	300	(272,	356)	459
Kennedy	205	(175,	217)	249
Roberts	190	(166,	214)	200
Alito	186	(165,	214)	203
Scalia	175	(157,	225)	65
Thomas	149	(133,	198)	9

Table 4.2: Conservatism rank (1-522) of each justice, joint model vs. JCS.

initial values described in section 4.4.1, the greater divergence of the liberal justices from each other than indicated by the JCS does not tell such a story, nor does the placement of each justice significantly closer to the center of the whole space than the party medians even though justice locations were initialized to the party means. Rather, the comparison of the scores estimated here with existing ones suggests that this model will be able to make a substantial new contribution to our understanding of how justices' ideology relates to that of other politicians when estimated using more data.

Results not related to justice and amicus votes largely conform with those of Bonica (2013). Even on this small set of data, the finding is largely reproduced that interest groups are more concentrated in the center of the ideological space than the candidates they support.

One point of interest is raised by the ideal point estimates for Presidential candidates. The estimates of location for individual Presidential candidates are better understood as representations of where they ran specifically in 2008 than as evaluations of their stable, long-term ideology. Candidates' success at the primary election stage, as expected, appears related to their running on positions toward the extremes of the ideological distribution. In particular, although Hillary Clinton is thought to have run well to the right of eventual nominee Barack Obama (whose position is not estimated, due to his use of public funding in lieu of PAC contributions), she also appears to have run substantially to the

left of other Democrats in terms of funding sources. Likewise, serious contenders John McCain and Rudy Giuliani appear to have run well to the right of Mitt Romney in 2008; it is not expected that an estimate of Romney's position on data from the 2012 election cycle would place him as close to the center of the political space. The wide variation in location among the seven Presidential candidates, the wide confidence intervals for each given data from a single election cycle, and the location of candidates such as Romney and Clinton in unexpected parts of the ideological space suggest that Presidential candidates who run more than once may do so in quite different ideological niches from campaign to campaign, in contrast to legislators whose ideologies are understood to be largely stable over time.

4.6 Conclusion

Amicus curiae briefs provide a bridge between models of judicial ideology based on justice votes and spatial models of ideology for organized interests and elected politicians based on campaign contributions from the former to the latter. By treating organized interests who also file amicus briefs on Supreme Court cases as sincere voters on those cases, and estimating a model for the votes of both amici and justices together with a spatial count model for contributions, we can compare the ideological locations revealed by justices through their votes to those of the interests themselves, and, therefore, to those of Congressional and Presidential incumbents and challengers. Such a joint model improves on previous bridging efforts such as Epstein et al. (2007) by greatly increasing the number of bridging observations used and permitting, in principle, unconditional estimates of the uncertainty in each parameter.

This chapter presents only a demonstration of the merit of this model, with much work remaining to be done. While the results presented in this chapter using data from a single election cycle establishes the feasibility of the method described, the joint model will be

of substantial usefulness to political science only when it provides ideal point estimates comparable across branches for multiple terms. The most important step forward will thus be to estimate the model on data from a longer period. With a longer period of time under consideration, it will become possible to check the model's performance by examining, for instance, how strongly ideological distance between a justice in their first Court term and a Senator at the time of their confirmation predicts whether the Senator voted to confirm the justice. Other obvious questions of interest will be made answerable with more years of data as well; for instance, it will be interesting to view how close the ideal points of the case parties in *Bush v. Gore* are estimated to lie to the locations of the actual presidential candidates who contested the case.

Improvements to the model itself as well as the data set are also possible. In conjunction with the expansion of the data set, it will be helpful to identify reasonable simplifications to reduce the number of parameters being estimated, especially the number of coefficients for non-spatial determinants of campaign contributions. Modifications to the model such as reparameterization or respecification for easier likelihood maximization and the re-introduction of an abstention option for amici should be tried, and such modifications should be tested quantitatively for whether they actually improve model fit. Lastly, while maximum likelihood estimation and the parametric bootstrap produce point estimates and confidence intervals of reasonable quality for demonstration, it will be best ultimately to overcome the obstacles to Markov chain Monte Carlo sampling of the model; MCMC estimates, if they can be produced, will not suffer from the kinds of bias described in section 4.4.2.

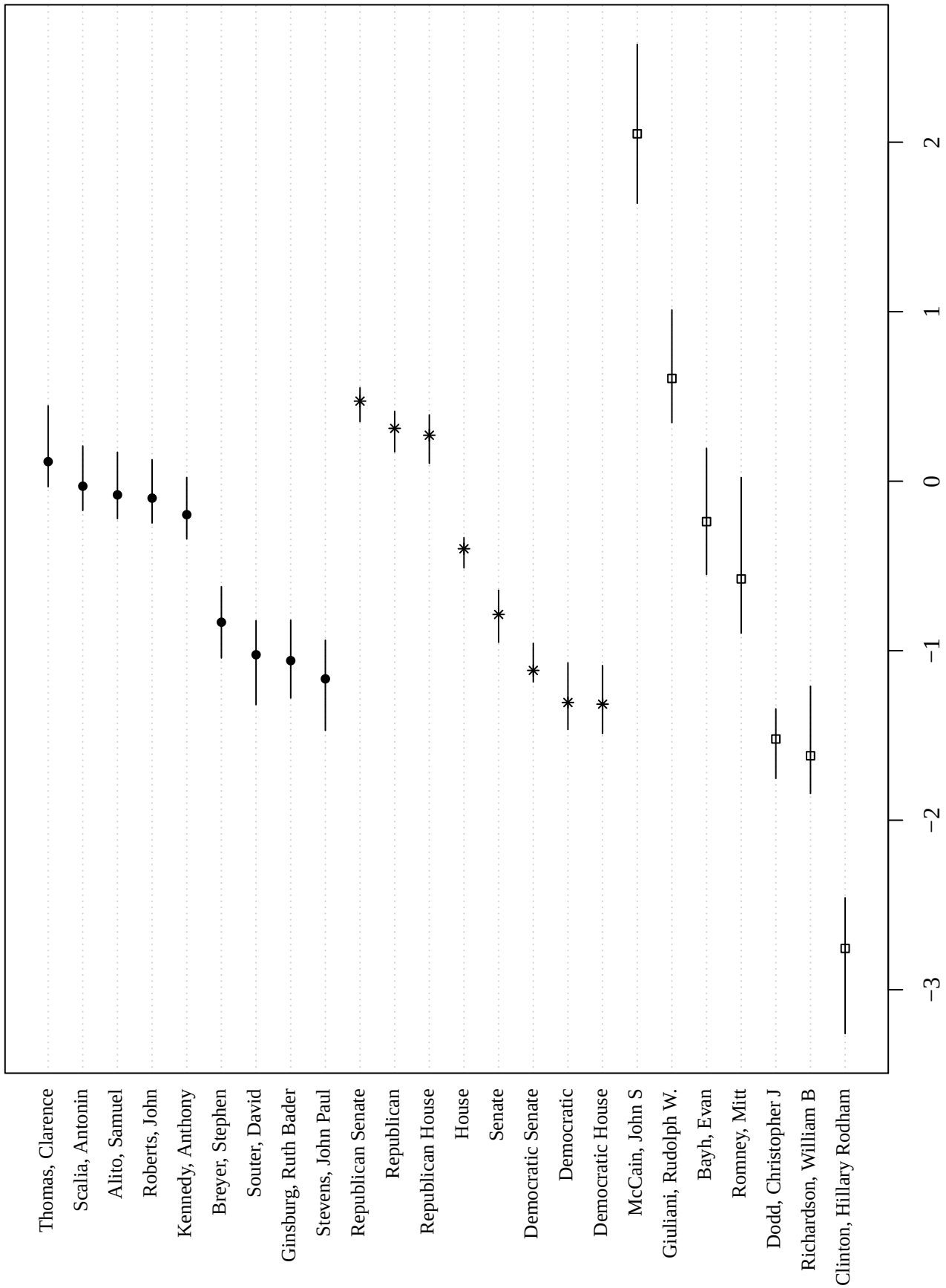

Figure 4.1: Ideal points of justices, median legislative candidates, and presidential candidates.

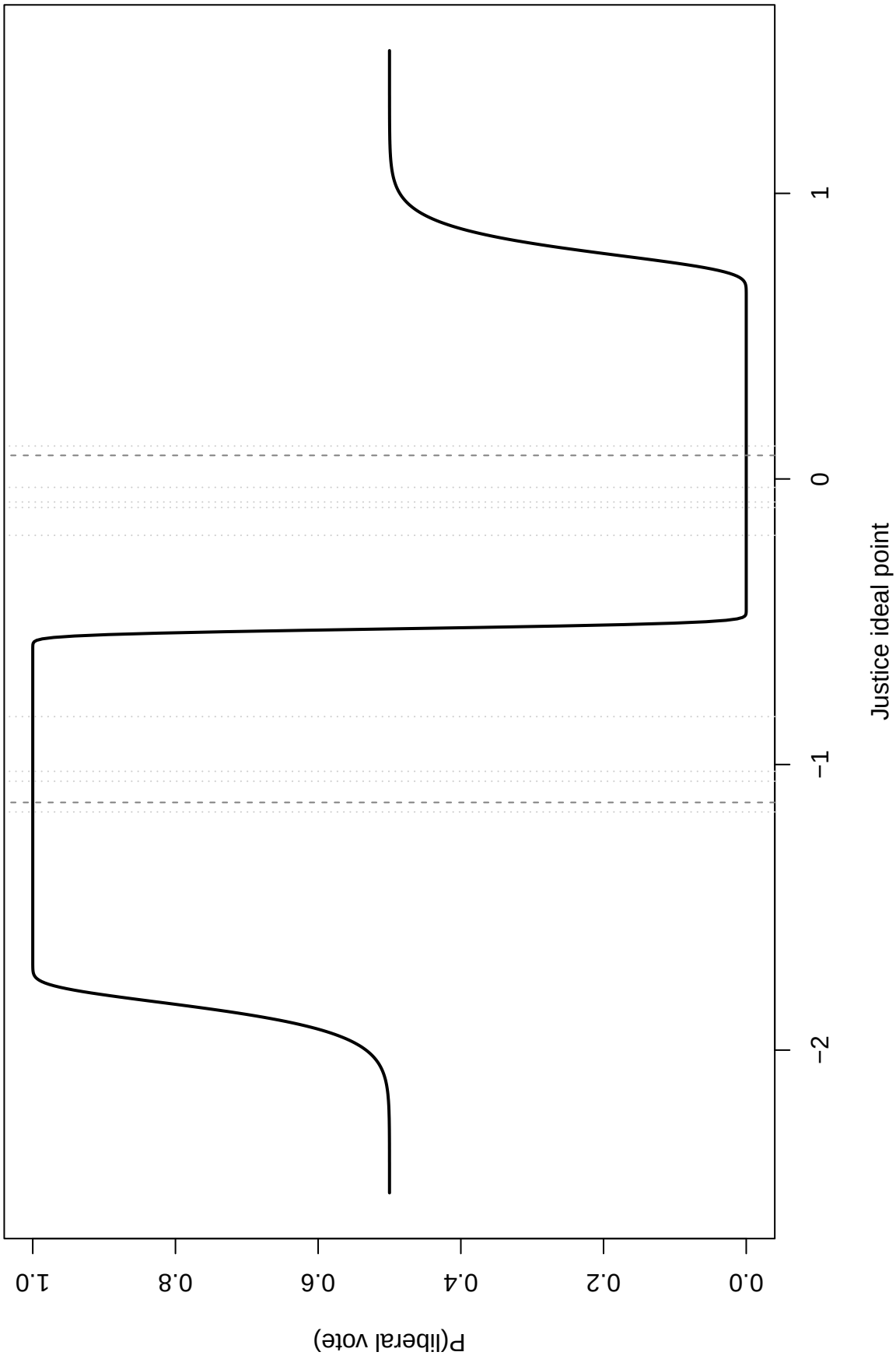

Figure 4.2: Probability of liberal vote vs. justice ideal point given median-located case parties.

Bibliography

Adelstein, Richard P. 1978. "The Plea Bargain in Theory: A Behavioral Model of the Negotiated Guilty Plea." *Southern Economic Journal* 490(3):488–503.

Bailey, Michael A. 2007. "Comparable preference estimates across time and institutions for the court, congress, and presidency." *American Journal of Political Science* 51(3):433–448.

Baker, Scott and Claudio Mezzetti. 2001. "Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial." *Journal of Law, Economics, and Organization* 17:149–167.

Bar-Gill, Oren and Oren Gazal Ayal. 2006. "Plea Bargains Only for the Guilty." *Journal of Law and Economics* 49(1):353–364.

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace." *American Journal of Political Science* 57(2):294–311.

Boylan, Richard T. 2005. "What Do Prosecutors Maximize? Evidence from the Careers of U.S. Attorneys." *American Law and Economics Review* 7:397–402.

Calvert, Randall L. 1995. *Explaining Social Institutions*. University of Michigan Press chapter Rational Actors, Equilibrium, and Social Institutions. Edited by Jack Knight and Itai Sened.

Campbell, W. K. and C. Sedekides. 1999. "Self-Threat Magnifies the Self-Serving Bias." *Review of General Psychology* 3:23–43.

Carroll, Royce, Jeffrey B Lewis, James Lo, Keith T Poole and Howard Rosenthal. 2009. "Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap." *Political Analysis* 17(3):261–275.

Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole and Howard Rosenthal. 2013. "The Structure of Utility in Spatial Models of Voting." *American Journal of Political Science* 57:1008–1028.

Carroll, Royce, Jeffrey B. Lewis, James Lo, Nolan McCarty, Keith Poole and Howard Rosenthal. 2013. ""Common Space" DW-NOMINATE Scores With Bootstrapped Standard Errors." World Wide Web page.
URL: *http://voteview.com/dwnomin_joint_house_and_senate.htm*

de Dreu, Carsten K. W., Bianca Beersma, Wolfgang Steinel and Gerben A. van Kleef. 2007. *Social Psychology: Handbook of Basic Principles*. 2 ed. The Guilford Press chapter The Psychology of Negotiation, pp. 608–639.

Dickson, Eric. 2006. "Rational Choice Epistemology and Belief Formation in Mass Politics." *Journal of Theoretical Politics* 18(4):454–497.

Diekmann, K. A., A. E. Tenbrunsel and A. D. Galinsky. 2003. "From Self-Prediction to Self-Defeat: Behavioral Forecasting, Self-Fulfilling Prophecies, and the Effect of Competitive Expectations." *Journal of Personality and Social Psychology* 85:672–683.

Efron, Bradley. 1987. "Better bootstrap confidence intervals." *Journal of the American statistical Association* 82(397):171–185.

Efron, Bradley and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Vol. 57 CRC press.

Ely, Jeffrey C. and Okan Yilankaya. 2001. "Nash Equilibrium and the Evolution of Preferences." *Journal of Economic Theory* 97:255–272.

Epstein, Lee, Andrew D Martin, Jeffrey A Segal and Chad Westerland. 2007. "The judicial common space." *Journal of Law, Economics, and Organization* 23(2):303–325.

Fey, Mark and Kristopher W. Ramsay. 2006. "The Common Priors Assumption: A Comment on "Bargaining and the Nature of War"." *Journal of Conflict Resolution* 50(4):607–613.

Filson, Darren and Suzanne Werner. 2002. "A Bargaining Model of War and Peace: Anticipating the Onset, Duration, and Outcome of War." *American Journal of Political Science* 46(4):819–838.

Frank, Robert H. 1988. *Passions within reason: The strategic role of the emotions.* WW Norton & Co.

Gill, Jeff and Gary King. 2004. "What to Do When Your Hessian is Not Invertible Alternatives to Model Respecification in Nonlinear Estimation." *Sociological methods & research* 33(1):54–87.

Gintis, Herbert. 2009. *Game Theory Evolving*. 2 ed. Princeton University Press.

Glaeser, Edward L., Daniel P. Kessler and Anne Morrison Piehl. 2000. "What do Prosecutors Maximize? An Analysis of the Federalization of Drug Crimes." *American Law and Economics Review* 2(2):259–290.

Gordon, Sanford C. and Gregory A. Huber. 2002. "Citizen Oversight and the Electoral Incentives of Criminal Prosecutors." *American Journal of Political Science* 46:334–351.

Gordon, Sanford C. and Gregory A. Huber. 2009. "The Political Economy of Prosecution." *Annual Review of Law and Social Science* 5:135–156.

Grossman, Gene M. and Michael L. Katz. 1983. "Plea Bargaining and Social Welfare." *The American Economic Review* 73(4):749–757.

Guth, W. and M. Yaari. 1992. *An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game*. University of Michigan Press.

Hansford, Thomas G. 2012. "Using the Amici Network to Measure the Ideological Loading of Supreme Court Cases." Working paper.
URL: *http://faculty.ucmerced.edu/thansford/Working%20Papers/Hansford_MPSA_2012.pdf*

Heifetz, Aviad, Chris Shannon and Yossi Spiegel. 2007*a*. "The Dynamic Evolution of Preferences." *Economic Theory* 32(2):251–286.

Heifetz, Aviad, Chris Shannon and Yossi Spiegel. 2007*b*. "What to Maximize If You Must." *Journal of Economic Theory* 133:31–57.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47:263–291.

Kahneman, Daniel and Amos Tversky. 1995. *Barriers to the Negotiated Resolution of onflict*. Norton chapter Conflict Resolution: A Cognitive Perspective, pp. 49–67.

Kessler, Daniel P. and Anne Morrison Piehl. 1998. "The Role of Discretion in the Criminal Justice System." *Journal of Law, Economics, and Organization* 14(2):256–276.

Kreps, David M. and Robert Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory* 27:253–279.

Landes, William M. 1971. "An Economic Analysis of the Courts." *Journal of Law and Economics* 14(1):61–107.

Lewis, Jeffrey B and Keith T Poole. 2004. "Measuring bias and uncertainty in ideal point estimates via the parametric bootstrap." *Political Analysis* 12(2):105–127.

Londregan, John. 1999. "Estimating legislators' preferred points." *Political Analysis* 8(1):35–56.

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999." *Political Analysis* 10(2):134–153.

Maynard Smith, J. and G. R. Price. 1973. "The Logic of Animal Conflict." *Nature* 246:15–18.

McCarthy, Davis J, Yunshun Chen and Gordon K Smyth. 2012. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic acids research* 40(10):4388–4297.

Oechssler, Jorg and Frank Riedel. 2001. "Evolutionary Dynamics on Infinite Strategy Spaces." *Economic Theory* 17(1):141–162.

Ok, Efe A. and Fernando Vega-Redondo. 2001. "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario." *Journal of Economic Theory* 97:231–254.

Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action." *American Political Science Review* 92(2):1–22. Presidential Address, American Political Science Association, 1997.

Paese, P. W. and R. D. Yonker. 2001. "Toward a Better Understanding of Egocentric Fairness Judgments in Negotiation." *International Journal of Conflict Management* 12:97–186.

Poole, Keith T. and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science* 29(2):357–384.

Powell, Robert. 2004. "Bargaining and Learning while Fighting." *American Journal of Political Science* 48(2):344–361.

Reinganum, Jennifer F. 1988. "Plea Bargaining and Prosecutorial Discretion." *Ame* 78:713–728.

Ross, L. and A. Ward. 1995. *Adcances in Experimental Social Psychology*. Vol. 27 Academic Press chapter Psychological Barriers to Dispute Resolution, pp. 255–304.

Rubin, J. Z., D. G. Pruitt and S. H. Kim. 1994. *Social Conflict: Escalation, Stalemate, and Settlement*. Academic Press.

Samuelson, Larry. 2001. "Introduction to the Evolution of Preferences." *Journal of Economic Theory* 97(2):225–230.

Selten, Reinhard. 1978. "The chain store paradox." *Theory and decision* 9(2):127–159.

Sethi, Rajiv and E. Somanathan. 2001. "Preference Evolution and Reciprocity." *Journal of Economic Theory* 97:273–297.

Slantchev, Branislav L. 2003. "The Principle of Convergence in Wartime Negotiations." *American Political Science Review* 97(4):621–632.

Smith, Alastair. 1988. "Fighting Battles, Winning Wars." *Journal of Conflict Resolution* 42:301–320.

Smith, Alastair and Allan C. Stam. 2004. "Bargaining and the Nature of War." *Journal of Conflict Resolution* 48(6):783–813.

Smith, Alastair and Allan C. Stam. 2006. "Divergent Beliefs in "Bargaining and the Nature of War": A Reply to Fey and Ramsay." *Journal of Conflict Resolution* 50(4):614–618.

Smith, John Maynard. 1982. *Evolution and the Theory of Games*. Cambridge University Press.

Tsebelis, George. 1990. "Are sanctions effective? A game-theoretic analysis." *Journal of Conflict Resolution* 34(1):3–28.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185:1124–1131.

Tversky, Amos and Daniel Kahneman. 1986. "Rational Choice and the Framing of Decisions." *The Journal of Business* 59(2):S251–S278.

Young, H. Peyton. 1993. "The Evolution of Conventions." *Econometrica* 61(1):57–84.