Preference Evolution, Two-Speed Dynamics, and Rapid Social Change[*]

William H. Sandholm
Department of Economics
University of Wisconsin
1180 Observatory Drive
Madison, WI 53706
whs@ssc.wisc.edu
www.ssc.wisc.edu/~whs

First Version: October 22, 1998
This Version: October 26, 2000

[*] I thank Kelly Chang, Eddie Dekel, Steven Durlauf, Werner Güth, Timur Kuran, Bart Lipman, Nolan Miller, Larry Samuelson, and Art Shneyerov, as well as seminar participants at Wisconsin, the 1999 Stony Brook Conference, and the 2000 American Economic Association Meeting for their comments.

Abstract

We present a dynamic analysis of the evolution of preferences in a strategic environment. In our model, each player’s behavior depends upon both the game’s payoffs and his idiosyncratic biases, but only the game’s payoffs determine his evolutionary success. Dynamics run at two speeds at once: while natural selection slowly reshapes the distribution of preferences, players quickly learn to behave as their preferences dictate. We establish the existence and uniqueness of the paired trajectories of society's preferences and aggregate behavior. While aggregate behavior adjusts smoothly in equilibration games, in coordination games aggregate behavior can jump discretely in an instant of evolutionary time.

1. Introduction

The origins of evolutionary game theory lie in biological models of natural selection. The players in these models are animals genetically programmed to play a certain strategy; evolution is driven by differences in their reproductive success. Economists have adapted models from evolutionary game theory to study the dynamics of human behavior. As the behavior of economic agents is driven by conscious choice rather than natural selection, the economic models describe how agents *learn* to satisfy their preferences.

While economists usually study the behavior of populations whose preferences are fixed, it is worthwhile to try to explain where these preferences come from. Unlike their behavior, the preferences of economic agents are traits determined by natural selection. Preferences which lead to reproductive success thrive at the expense of the others. Of course, the effects of preferences on reproductive fitness are mediated through behavior: preferences determine an agent's actions, which in turn determine his fitness. Because preferences generate fitness in this indirect way, to understand preference evolution one must act as biologist and economist at once. While one shows how the distribution of preferences is shaped by natural selection, one must concurrently describe how agents learn to behave as their preferences dictate.

Although learning and natural selection occur simultaneously, the former proceeds much more quickly than the latter: players can quickly switch to preferred strategies, but changes in the distribution of preferences are driven by gradual turnover in the population's membership. Preference evolution therefore leads us to consider dynamics which run at two speeds at once. We shall see that these dynamics require special techniques of analysis, and that they define a model of preference evolution with features not present in models of learning or natural selection alone.

In this paper, we study the evolution of preferences in a strategic setting. In our model, a single population of players repeatedly plays a two strategy game, the payoffs of which represent evolutionary fitnesses. Each player’s utility function combines the common fitness function with his idiosyncratic biases. While preferences determine individual behavior, society's aggregate behavior determines which actions are fit, and hence which preferences survive. As the preference distribution is shaped by natural selection, behavior adjusts in tandem, as players alter their strategy choices to maintain equilibrium play.

We establish the existence and uniqueness of the solution trajectories of this evolutionary process. We find that in equilibration games (e.g., the Hawk-Dove game), aggregate behavior adjusts continuously in response to changes in the distribution of preferences. In contrast, strategy adjustment in coordination games may be discontinuous: discrete changes in the overall strategy distribution can occur in an instant of evolutionary time. Thus, when society must learn to distribute itself between strategies, the adjustment process is smooth; when society must agree upon a convention, consensus may emerge in leaps and bounds.

We study the evolution of biases, by which we mean idiosyncratic preferences for certain modes of behavior. In our model, biases are simply predispositions towards or against each strategy; a player's overall payoff to choosing a strategy sums the strategy's fitness and his personal bias. Since each player's behavior only depends on the difference between the two strategies' payoffs, we summarize his biases by a single number reflecting his relative bias towards strategy A.

Holding its biases fixed, a population’s behavior is in equilibrium if no player can unilaterally improve his payoffs. Every equilibrium can be characterized by a bias threshold, θ: players whose (relative) bias towards strategy A exceeds θ choose strategy A, while the others choose strategy B. In the equilibrium, a player whose bias is exactly θ is indifferent between the two strategies.

While players’ decisions are influenced by their bias, their evolutionary prospects are not. Thus, differences in fitness may exist even when players’ behavior is in equilibrium.[1] Suppose that strategy A yields a higher expected fitness than strategy B. Then players who choose A are more fit than the others. Therefore, if behavior is described by the bias threshold θ, then biases above θ (i.e., those which prompt play of strategy A) will become more prevalent at the expense of those below. Since players can switch strategies much faster than the population's preference composition can change, we assume that as preferences evolve, players instantly adjust their behavior to maintain equilibrium play.

We first use this model to study *equilibration games*, which are games in which increasing the number of players choosing a strategy lowers its relative fitness, and in which neither strategy is dominant.[2] We fix an arbitrary initial distribution of biases and ask how both the bias distribution and the population's aggregate

[1] For example, even if strategy A is dominant in the underlying game, strategy B may be played in equilibrium by players who are sufficiently biased towards it.

[2] Examples of equilibration games include the Hawk-Dove game and games modeling choice between goods exhibiting negative consumption externalities.

behavior change in response to evolutionary pressures. We find that as the bias distribution evolves, aggregate behavior adjusts continuously to maintain equilibrium play.

Behavior trajectories can look quite different in *coordination games*: games in which increasing the number of players using a strategy increase its relative fitness, and in which neither strategy is dominant.[3] In these games, preference evolution often forces aggregate behavior to adjust discontinuously: as the distribution of preferences changes, a moment is reached at which equilibrium play can only be maintained if a significant fraction of the players simultaneously switch strategies. Thus, when a population benefits from acting in concert, we should expect sudden, seemingly unprovoked shifts in the way it behaves.

We can provide intuition for these results by decomposing the influence of preference evolution on aggregate behavior into two distinct effects. As an example, consider a situation in which strategy *A* yields higher fitness than strategy *B*. In this case, evolution causes biases above the threshold (i.e., those which prompt strategy *A*) to become more prevalent in the population. The primary effect of this change in preferences is to increase the proportion of players who choose strategy *A*.

As preferences change, players adjust their behavior to maintain equilibrium play. If the underlying contest is an equilibration game, then increasing the representation of strategy *A* makes this strategy less attractive relative to strategy *B*. Thus, the secondary effect of the good performance of strategy *A* is that players who had marginally preferred to play *A* will begin to play *B*. This secondary effect *inhibits* the growth of strategy *A*.

On the other hand, if the players face a coordination game, then increasing the representation of strategy *A* makes this strategy more attractive. As biases prompting strategy *A* become more prevalent, nearly indifferent players switch from strategy *B* to strategy *A*, *reinforcing* the growth of strategy *A*. Surprisingly, this secondary effect can dominate the primary effect, even when the number of indifferent players is moderate. Indeed, the secondary effect can lead behavior to adjust at an infinite rate: we show that if the density of the indifferent players approaches a finite bound determined by the underlying fitnesses, only a discrete change in aggregate behavior is enough to preserve equilibrium play.

[3] It will become clear after we introduce our formal model that this definition generalizes most standard definitions of two strategy coordination games.

Coordination games can admit many equilibrium behaviors under a single, fixed bias distribution.[4] In order to single out the equilibria which are robust to small changes in behavior, we consider how a population whose preferences are fixed adjusts its behavior to achieve equilibrium play. When studying preference evolution, we assume that players always follow an equilibrium which is stable with respect to this adjustment process. Doing so ensures that the jumps in our model do not arise because an unstable equilibrium is disrupted by small behavior trembles. Rather, preference evolution causes stable equilibria to suddenly vanish, forcing rapid adjustment to a new, stable behavior configuration.

Our model allows players' biases to take on a continuum of values; it is therefore a continuous time dynamical system with an infinite dimensional state space. It exhibits two types of discontinuities: discontinuities of preference growth rates, which depend on a discrete strategy choice, and the jumps in behavior described above. The main technical results of this paper establish the existence and uniqueness of solution trajectories to this apparently complicated dynamical system.

Güth and Yaari (1992) and Güth (1995) investigate preference evolution in a resource acquisition game. Using a static notion of preference stability, they show how preferences for punishing cheaters can be evolutionarily stable, thereby providing an evolutionary explanation for cooperative behavior. However, their results rely heavily on the assumption that players' preferences are sometimes commonly known. In contrast, we assume that players can only learn the aggregate distribution of behavior; no knowledge of opponents' preferences is ever required.[5]

Huck and Oechssler (1999) study the evolution of preferences for rejecting greedy proposals in the Divide the Dollar game. They assume that preferences are unobservable, so that players' choices only depend on the population's aggregate behavior. Huck and Oechssler (1999) show that when players interact in small groups, evolution favors preferences for rejection, rendering fair division the unique evolutionarily stable outcome.

Ely and Yilankaya (1997) define a notion of the evolutionary stability of preference distributions for populations playing normal form games. In their model, all individuals have expected utility preferences over the set of strategy

[4] For example, if most players are nearly unbiased, there are three equilibria, one approximating each of the equilibria (two pure, one mixed) of the underlying game. We shall see that for any given coordination game, there are bias distributions which generate any specified number of equilibria.

[5] For further results on the stability of preferences when preferences are observable, see Dekel, Ely, and Yilankaya (1998) and Huck, Kirchsteiger, and Oechssler (1999).

profiles, but unlike in our own model, no connection between these preferences and the underlying fitnesses is assumed. Ely and Yilankaya (1997) prove the existence of stable preference/behavior pairs in all normal form games. However, the strength of their result is mitigated by the weakness of the restriction they place on the evolutionary stability of behavior when preferences are fixed. We avoid this difficulty by restricting attention to equilibria which are robust to the slight behavior disturbances which population turnover inevitably creates.[6]

We should emphasize the central difference between the models just described and our own model. All of the other models focus on the preference/behavior pairs which constitute stable equilibria. In contrast, our primary concern is with the out-of-equilibrium dynamics which arise when preferences and behavior simultaneously evolve. Our main results show that continuous preference evolution can generate discontinuous shifts between stable behavior configurations, but only when the players face a coordination game.

Our model is closely related to work of Kuran (1989, 1991, 1995) on political and cultural revolutions. Kuran models conflicts between political and cultural regimes using a form of coordination game in which players have idiosyncratic preferences. Under certain conditions, small changes in preferences can lead to large and rapid shifts in behavior. Using this model, Kuran explains the sudden and surprising nature of the 1978-79 revolution in Iran and the 1989 revolution in Eastern Europe as consequences of small, exogenous shifts in preferences. We study behavior in abstract coordination games with diverse preferences, and we introduce an explicit mechanism through which behavior endogenously influences preferences. An application of this mechanism in the contexts considered by Kuran may provide an endogenous explanation for rapid shifts between political regimes.[7]

The remainder of the paper proceeds as follows. Section 2 defines fitness and bias. Section 3 introduces the model of preference evolution, which is used in Section 4 to study equilibration games. Section 5 analyzes behavior adjustment

[6] A number of papers (Karni and Schmeidler (1986), Cooper (1987), Robson (1996a, 1996b), and Dekel and Scotchmer (1999)) investigate the biological foundations of risk preferences. Of these, only Robson (1996b) and Dekel and Scotchmer (1999) consider biological evolution in a strategic setting. Both papers show how payoff discontinuities which are inherent in the competition among males for mates may provide an evolutionary basis for preferences for risk-seeking behavior.

[7] Our model of preference evolution and rapid social change in games is closely related to models from catastrophe theory (see, e.g., Poston and Stewart (1977)). More precisely, our model can be regarded as one of catastrophes generated by feedback from a one-dimensional behavior space into an infinite-dimensional control space. However, as the tools of catastrophe theory do not appear to provide additional insights into our model, we will not pursue this connection here.

under a fixed preference distribution. Section 6 incorporates this analysis into an extended model of preference evolution used to study coordination games. Section 7 discusses the connections between our model and Harsanyi's (1973) model of the purification of mixed strategy equilibria. Section 8 concludes.

2. Basic Definitions

2.1 Fitness

A continuum of players repeatedly plays a two-strategy game. Each player's payoffs are the sum of two components: the *fitness*, which depends on the realized strategy distribution and the player's action, and the *bias*, which depends solely on the player's action and which varies idiosyncratically from player to player. While both fitness and bias determine players' behavior, only fitness leads to evolutionary success. In this section, we consider fitnesses, which we express as the payoffs of the underlying game. Biases will be introduced in Section 2.2.

Players repeatedly play the game $G = \{\{A, B\}, \{\phi_A, \phi_B\}\}$. The C^2 functions ϕ_A: $[0, 1] \rightarrow \mathbf{R}$ and ϕ_B: $[0, 1] \rightarrow \mathbf{R}$ represent the fitnesses of strategies A and B given the proportion of players $x \in [0, 1]$ choosing strategy A. In the biological interpretation of our model, fitness is a measure of a strategy's reproductive success.[8]

We call $L(x) = \phi_A'(x) - \phi_B'(x)$ the *alignment* of the game. The alignment measures the degree to which acting in concert benefits the players: when aggregate behavior is x, $L(x)$ measures the marginal change in each strategy's relative fitness when the number of players choosing that strategy increases.[9]

We restrict attention to games in which the sign of $L(x)$ is the same for all $x \in [0, 1]$, and in which neither strategy is dominant.[10] We call G an *equilibration game* if it has negative alignment and no dominant strategy. Such games possess a unique equilibrium, $x^* \in (0, 1)$, which satisfies $\phi_A(x^*) = \phi_B(x^*)$. Game G is a *coordination*

[8] In certain economic applications, we can instead interpret fitness as some other objective (e.g., financial) measure of success: see Section 3.3.

[9] That is, the rate of improvement in strategy A's relative payoffs when the number of players choosing strategy A increases is equal to $\frac{d}{dx}(\phi_A(x) - \phi_B(x)) = L(x)$, while the rate of improvement in strategy B's relative payoffs when the number of players choosing strategy B increases is equal to $\frac{d}{d(1-x)}(\phi_B(x) - \phi_A(x)) = \frac{d}{dx}(\phi_A(x) - \phi_B(x)) = L(x)$.

[10] Strategy S is *dominant* if $\phi_S(x) \geq \phi_{S'}(x)$ for $S' \neq S$ and for all $x \in [0, 1]$. Our results are easily extended to games with a dominant strategy.

game if it has positive alignment and no dominant strategy. Coordination games have three equilibria: one, $x^* \in (0, 1)$, in which both strategies are used and are equally fit, and two in which all players coordinate on the same strategy.

As an example, suppose that players are randomly matched to play the 2 x 2 symmetric game illustrated in Figure 1.

	A	B
A	a, a	b, c
B	c, b	d, d

Figure 1: A 2 x 2 symmetric game.

If $x \in [0, 1]$ is the proportion of players in the population choosing strategy A, then the (expected) fitness of players using strategies A and B are

$$\phi_A(x) = ax + b(1-x);$$
$$\phi_B(x) = cx + d(1-x).$$

Since the game's payoffs are linear in x, its alignment is constant: $L(x) = \phi_A'(x) - \phi_B'(x) = a - b - c + d$. We can therefore partition random matching games into three classes which are the linear cases of the classes defined above: coordination games ($a > c$, $b < d$), equilibration games ($a < c$, $b > d$), and games with a dominant strategy ($a \geq c$, $b \geq d$, or $a \leq c$, $b \leq d$). In the first two classes of games, the mixed equilibrium puts mass $x^* = \frac{d-b}{a-b-c+d}$ on strategy A.

2.2 Bias

In most settings, we expect players' preferences over outcomes to be related to the fitnesses of those outcomes; still, other factors may influence the preferences' ultimate form. A player's own strategies, the elements of his choice set, are the most salient aspects of his strategic environment. It therefore seems natural for idiosyncracies in preferences to take the form of biases towards each strategy. We assume that each player's payoff functions sum the fitness functions and his biases. While a player's biases influence how he behaves, only his fitness determines his reproductive success.

Each player's biases are described by a pair of scalars, α and β, which represent his unobservable, idiosyncratic payoffs to playing strategy A and strategy B, respectively. His total payoffs are therefore given by

$$\pi_A(x) = \phi_A(x) + \alpha;$$
$$\pi_B(x) = \phi_B(x) + \beta.$$

Since players' choices only depend on the difference between $\pi_A(x)$ and $\pi_B(x)$, we can assume without loss of generality that β is zero. We therefore call α the player's *bias*. The distribution of biases in the population is described by a density function f: $\mathbf{R} \to \mathbf{R}_+$. We let F: $\mathbf{R} \to [0, 1]$ denote the corresponding *decumulative* distribution: $F(\alpha) = \int_\alpha^\infty f(\lambda)\,d\lambda$ for all $\alpha \in \mathbf{R}$.

For any fixed game and preference distribution, all equilibrium behaviors can be characterized by a *bias threshold*, $\theta \in \mathbf{R}$. When θ is the bias threshold, players whose bias towards A is at least θ choose A, while the remaining players choose B. For θ to represent an equilibrium, players whose bias is θ must be indifferent between strategies. We let $\sigma(\alpha, \theta)$ denote the strategy used by players whose bias is α when the threshold is θ:[11]

$$\sigma(\alpha, \theta) = \begin{cases} A & \text{if } \alpha \geq \theta, \\ B & \text{if } \alpha < \theta. \end{cases}$$

Observe that $x = F(\theta) = \int_\theta^\infty f(\lambda)\,d\lambda$ is the proportion of players who choose strategy A when the bias threshold is θ.

Define the function I: $[0, 1] \to \mathbf{R}$ by $I(x) = \phi_B(x) - \phi_A(x)$. $I(x)$ equals the fitness advantage of strategy B at strategy distribution x. More importantly, since $\phi_A(x) + I(x) = \phi_B(x)$, we can interpret $I(x)$ as the bias which generates indifference between strategies A and B when aggregate behavior is given by x. We therefore call $I(\cdot)$ the *indifference function* associated with the game G.

Suppose that play is described by a bias threshold which is a fixed point of $I(F(\cdot))$: that is, $\theta = I(F(\theta))$. Since proportion $F(\theta)$ of the players choose strategy A, only the players with bias $I(F(\theta))$ are indifferent between strategies: players with higher biases strictly prefer A, while players with lower biases strictly prefer B. Since the bias

[11] Our assumption that players on the threshold choose strategy A is arbitrary; making alternative assumptions would not affect our results.

threshold θ equals $I(F(\theta))$, these are precisely the players choosing strategies A and B, respectively. Hence, if $\theta = I(F(\theta))$, θ is an equilibrium. Conversely, if $\theta \neq I(F(\theta))$, players whose biases lie between these two values would prefer to switch strategies. We therefore define

$$E_G(f) = \{\theta\text{:}\;\; \theta = I(F(\theta))\}.$$

to be the set of equilibria for the game G under bias distribution f.

Figure 2 illustrates the (inverse) indifference function I^{-1} of an equilibration game with linear payoffs. Since $I'(x) = -L(x)$, and since equilibration games have negative alignment, the indifference function and its inverse are both increasing. By the definition of x^*, $I(x^*) = \phi_B(x^*) - \phi_A(x^*) = 0$, so $I^{-1}(0) = x^*$. Figure 2 also contains

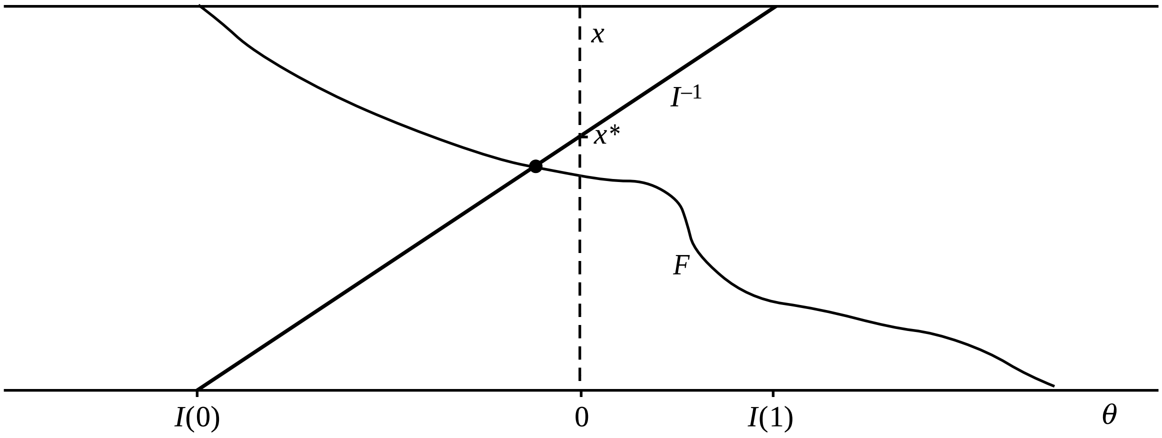

Figure 2: Equilibrium in an equilibration game.

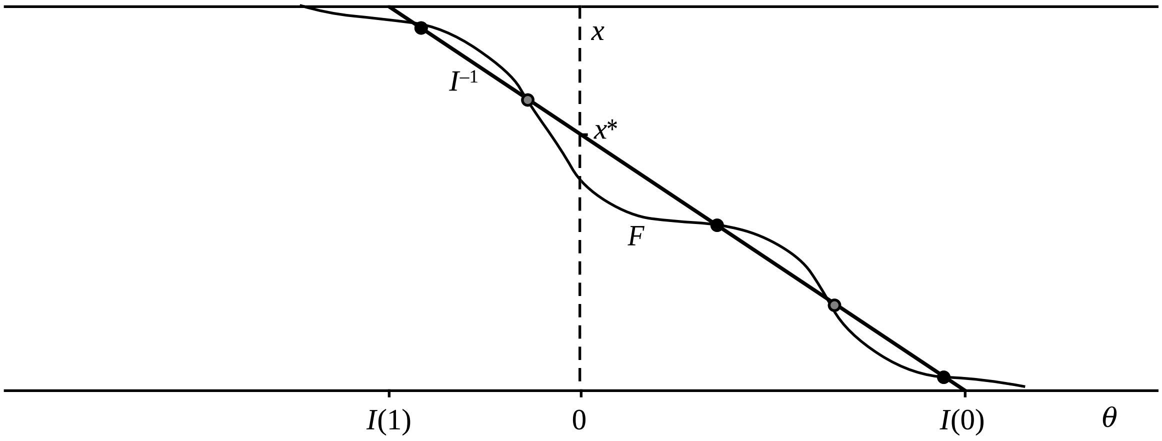

Figure 3: Equilibria in a coordination game.

a decumulative bias distribution, F. Equilibria are represented by the points at which the two curves intersect. Since F is decreasing, it crosses I^{-1} exactly once: equilibrium is unique.

Figure 3 contains the indifference function of a coordination game with linear payoffs. Coordination games have positive alignment, so this function is decreasing. Since F is also decreasing, I^{-1} and F can intersect many times: multiple equilibria are possible.

3. Evolutionary Dynamics

In this section we introduce our model's evolutionary dynamics. We begin by presenting the standard model of evolution in games, in which each strategy's fitness directly determines the change in its representation. Using this model as a point of departure, we introduce our model of preference evolution, in which preferences replace strategies as the object of evolutionary selection.

3.1 The Standard Model of Evolution

Let $x \in [0, 1]$ equal the proportion of players currently choosing strategy A. Evolution in the standard model is governed by the equation

$$\text{(S)} \qquad \dot{x}_t = x_t \, g_A(x_t),$$

where the C^1 functions g_A: $[0, 1] \to \mathbf{R}$ and g_B: $[0, 1] \to \mathbf{R}$ denote the percentage growth rates of strategies A and B as functions of the current strategy distribution. Since greater fitness yields greater reproductive success, the percentage growth rates satisfy

$$\text{sgn}(g_A(x)) = \text{sgn}(\phi_A(x) - \phi_B(x)) \text{ for all } x \in (0, 1).$$

We must also restrict the growth rates to keep the total population mass constant:

$$x\, g_A(x) + (1 - x)\, g_B(x) = 0 \text{ for all } x \in [0, 1].$$

This condition implies that $g_A(1) = g_B(0) = 0$.

The analysis of the standard model is quite simple.[12] In equilibration games, evolution from all interior population states leads to the unique equilibrium x^*. In coordination games, evolution from all initial conditions other than the mixed equilibrium lead to one of the pure equilibria of the game; which equilibrium is reached is determined by the side of the mixed equilibrium on which play begins.

3.2 Preference Evolution

In our model of preference evolution, the pair (f_t, θ_t) represents the bias density and bias threshold at time $t \in \mathbf{R}_+$. Given this pair, we let $x_t = F_t(\theta_t) = \int_{\theta_t}^{\infty} f_t(\alpha)d\alpha$ denote the mass of players choosing strategy A. Our basic model of preference evolution consists of two equations: one describing changes in the preference distribution, and the other governing the population's behavior.

(D) For all $\alpha \in \mathbf{R}$, $\frac{d}{dt^+} f_t(\alpha) = f_t(\alpha)\, g_{\sigma(\alpha,\theta_t)}(x_t)$ and $f_t(\alpha)$ is continuous in t.

(E) $\theta_t \in E_G(f_t)$ for all t.

Earlier, we defined the functions g_A and g_B as the percentage growth rates of strategies in the standard evolutionary model. In condition (D), g_A and g_B are used to define the growth rates of preferences. The condition states that the percentage growth rate of each bias α depends only on the fitness of the strategy choice which it induces. All biases leading to the same behavior exhibit the same percentage growth rate.[13]

Condition (E) requires that at each moment in time, the population's behavior is in equilibrium given the current bias distribution. Implicitly, this condition requires that players instantly modify their behavior in response to changes in the distribution of preferences. It therefore manifests our intuition that changes in preferences develop much more slowly than changes in behavior.

Condition (E) stands in for an explicit description of the players' learning process. We shall see that in equilibration games, the unique equilibrium is always stable with respect to this process, so explicitly introducing the learning process would not

[12] For details see, e.g., Weibull (1995).

[13] Technical aside: At any point in time at which players with bias α switch strategies, the growth rate of this bias changes discontinuously. It is for this reason that the evolutionary dynamics must be defined using right hand derivatives. This in turn forces us to assume directly that each $f_t(\alpha)$ changes continuously over time.

alter our results. On the other hand, coordination games admit both stable and unstable equilibria. Before studying preference evolution in these games, we will present our model of behavior adjustment; we will then restrict attention to equilibria which are stable under this adjustment process.

Condition (E) also requires that behavior adjust arbitrarily more quickly than preferences evolve. While running the two processes at different rates is natural, allowing the relative rates of adjustment to be infinite may seem inappropriate. Fortunately, relaxing this assumption has only a minor impact on our conclusions. In the Appendix, we offer a model in which the relative rates of behavior adjustment and preference evolution are bounded, and show that our main model is the limiting version of the bounded rate model when the bound approaches infinity.

Finally, we should note that the standard model of evolution can be viewed as a special case of our model of preference evolution. If no biases between $I(0)$ and $I(1)$ receive mass in the initial distribution, then all players in the population regard either strategy A or strategy B as dominant. Because no player conditions his choices on the population's aggregate behavior, evolution of preferences reduces to evolution of strategies.

3.3 Interpretations

In standard models of evolution in games, strategies are the units of selection: changes in a strategy's representation depend directly on its performance. Here, preferences are the unit of selection; since preferences are rules for choosing strategies, the differential survival of preferences is mediated through this choice.[14] In the basic biological interpretation of our model, preferences which induce the behavior that is currently more fit reproduce more successfully than those which do not.

Our model can also be interpreted as one of the cultural transmission of preferences, as might be used to study political or cultural change.[15] Suppose that preferences which have lead to economic success are more likely to be passed down to subsequent generations than those which have not. Then while preferences

[14] If the players in the interaction are humans, the meaning of "preferences" is clear. If the players are other sorts of animals, preferences can represent biological mechanisms determining whether an animal follows behavior A or behavior B in response to the current behavioral environment.

[15] For a related discussion of preference formation, see Bowles (1998, Sections 2 and 3). For analyses of shifts between political and cultural regimes, see Kuran (1989, 1991, 1995).

determine behavior, the survival of preferences only depends on their economic success, here identified with fitness.

Finally, our model can be applied in certain settings in which the distribution of preferences never changes. For example, suppose that economic success dictates interaction frequency. Then even if the numbers of players with each preference never changes, preferences leading to economic success will be encountered more often as time passes. To model this phenomenon, we interpret $f(\alpha)$ as a measure of the interaction frequency of preference α rather than as the number of players with that preference. Fitness is a function of $x = \int_{\theta}^{\infty} f(\alpha)d\alpha$, the total interaction frequency of players choosing strategy A. Preferences leading to fit actions become more prevalent in the interactions.

4. Equilibration Games

Before starting our analysis, we impose three assumptions on the model's initial conditions.

(A1) $\theta_0 \in E_G(f_0)$;

(A2) $F_0(I(0)), F_0(I(1)) \in (0, 1)$;

(A3) f_0 is C^1.

Assumption (A1) requires that the population's initial behavior is an equilibrium given the initial preference distribution. Assumption (A2) requires that strategy A and strategy B are each dominant for some non-negligible set of players. Results for initial preference distributions which violate assumption (A2) are straightforward extensions of our stated results. Lastly, assumption (A3) asks that the initial preference distribution be continuously differentiable. This condition is used to prove that solution trajectories exist.

Theorem 1 characterizes the evolution of preferences in equilibration games.

Theorem 1: *Suppose G is an equilibration game, and let the initial condition* (f_0, θ_0) *satisfy assumptions* (A1) - (A3). *Then*:

(*i*: *Existence*) *There exists a unique solution trajectory,* $\{f_t, \theta_t\}_{t\geq 0}$, *to* (D) *and* (E).

(*ii*: *Continuity*) θ_t *and* x_t *change continuously over time.*

(*iii. Limit Behavior*) $\lim_{t\to\infty}\theta_t = 0$ *and* $\lim_{t\to\infty}x_t = x^*$.

Standard models of evolution focus directly on how the strategy distribution changes over time. If the number of strategies (and hence the dimension of the state space) is finite, and if the laws of motion are smooth, existence and uniqueness of solutions follows from standard results on ODEs. Our model of preference evolution satisfies neither of these properties: the preference distribution, f_t, is infinite dimensional object; the law of motion (D) is discontinuous at the bias threshold. In the Appendix, we demonstrate the existence and uniqueness of solutions by establishing a one-to-one correspondence between solutions to the dynamical system defined by equations (D) and (E) and solutions to a certain two-dimensional ODE (Theorem A3). Part (*i*) of Theorem 1, which guarantees the existence and uniqueness of solutions to (D) and (E), then follows from standard results.

Part (*ii*) of the theorem states that the bias threshold and aggregate behavior change continuously over time. Part (*iii*) characterizes limit behavior. As time tends to infinity, the population's behavior approaches the unique equilibrium of the fitness game, while the bias threshold approaches zero: in the limit, a player chooses strategy *A* if and only if he is biased towards *A*.

To explain the dynamics in greater detail, we draw an example in which $x_0 < x^*$ in Figure 4. Since *G* is a equilibration game, its alignment is negative, and its indifference function I^{-1} is increasing. At each instant *t*, the game's unique equilibrium, θ_t, is defined by the intersection of the indifference line and the decumulative distribution F_t.

Because *G* is a equilibration game, whenever fewer than x^* players are choosing strategy *A*, its fitness is higher than that of strategy *B*: $\phi_A(x) > \phi_B(x)$ whenever $x < x^*$. Therefore, equation (D) implies that all biases above the threshold θ_t become more prevalent and that all others languish. The distribution becomes steeper to the right of θ_t and shallower to the left of θ_t, and therefore shifts upward.

Since players whose bias exceeds the threshold play strategy *A*, the primary effect of the evolution of preferences on behavior is to increase the number of players choosing *A*. We can compute this primary effect as

$$\int_{\theta_t}^{\infty}\dot{f}_t(\alpha)\,d\alpha = \int_{\theta_t}^{\infty} g_A(x_t)f_t(\alpha)\,d\alpha = x_t\,g_A(x_t).$$

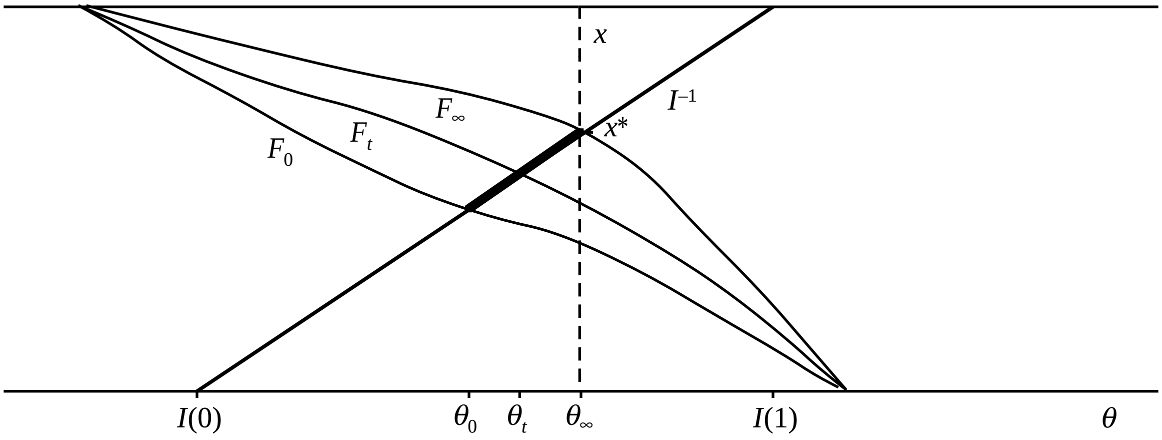

Figure 4: Evolution of preferences in an equilibration game.

Viewed in isolation, the primary effect is identical to the law of motion of the standard evolutionary model.

Preference evolution also gives rise to a secondary effect on behavior. As the preference distribution changes, equation (E) requires that the bias threshold adjust to maintain equilibrium play. Since the primary effect makes strategy *A* more common, it also makes strategy *A* less attractive: because *G* is an equilibration game, as *x* increases the fitness advantage of strategy *A* over strategy *B* falls. Therefore, players who had been indifferent between the two strategies begin to strictly prefer strategy *B*. To maintain equilibrium, the bias threshold θ_t must increase.

This secondary effect makes behavior change more slowly than in the standard model. To show this formally, we express aggregate behavior in two ways:

$$x_t \equiv \int_{\theta_t}^{\infty} f_t(\lambda)\, d\lambda;$$
$$x_t \equiv I^{-1}(\theta_t).$$

The first expression is the definition of x_t; the second is a form of the equilibrium condition (E). Because we have established that the solution trajectory $\{f_t,\ \theta_t\}_{t\geq 0}$ exists, we can differentiate these identities with respect to time. Let

$$\Lambda(\theta_t) \equiv L(I^{-1}(\theta_t)).$$

represent the game's alignment as a function of the bias threshold θ_t. Then differentiating yields[16]

$$\dot{x}_t = -f_t(\theta_t)\dot{\theta}_t + \int_{\theta_t}^{\infty} \dot{f}_t(\lambda)\,d\lambda = -f_t(\theta_t)\dot{\theta}_t + x_t g_A(x_t);$$
$$\dot{x}_t = -\Lambda(\theta_t)^{-1}\dot{\theta}_t.$$

Solving for $\dot{x}_t$, we obtain

$$\text{(B)} \qquad \dot{x}_t = \frac{x_t\; g_A(x_t)}{1 - \Lambda(\theta_t) f_t(\theta_t)}.$$

Since $\Lambda(\theta_t) < 0$, the denominator of this expression exceeds one. Thus, the speed of strategy adjustment is slower than that in the standard model. The greater the preference density at the threshold, $f_t(\theta_t)$, the more players who switch strategies at time t, and the more potent the dampening of the strategy adjustment process. Similarly, the greater the magnitude of the alignment, $\Lambda(\theta_t)$, the quicker the advantage of playing the “rarer” strategy falls as behavior equilibrates, and the slower aggregate behavior adjusts.

While t is finite, the mass of players choosing strategy A remains below x^*. Hence, strategy A remains more fit than strategy B, propelling further preference evolution. Only as time approaches infinity does the preference distribution settle. The support of the limit distribution is the same as that of the initial distribution: no biases become extinct during evolution’s course. Aggregate behavior converges to the equilibrium of the underlying game, eliminating the difference in fitness between the two strategies. Finally, the bias threshold approaches zero: in the limit, each player acts in direct accordance with his bias; preference evolution guarantees that this results in equilibrium play.

5. Behavior Adjustment

Condition (E) requires that as preferences evolve, the population maintains equilibrium play. Equilibration games have a unique equilibrium for each preference distribution, so condition (E) completely specifies the population's behavior at each moment in time. Since coordination games can exhibit multiple

[16] Since $I'(x_t) = -L(x_t)$, $(I^{-1})'(\theta_t) = I'(x_t)^{-1} = -L(x_t)^{-1} = -L(I^{-1}(\theta_t))^{-1} = -\Lambda(\theta_t)^{-1}$.

equilibria, analyzing these games requires more precise restrictions on behavior. We now motivate these restrictions by considering how a population whose preferences are fixed adjusts its behavior to establish equilibrium play.

Fix the preference distribution F. Our model of behavior adjustment assumes that the population's strategy choices can always be summarized by a bias threshold, θ.[17] Behavior adjustment is described by the equation

$$\dot{\theta} = h(\theta),$$

where the Lipschitz continuous function h satisfies

$$\text{(A)} \qquad \mathrm{sgn}(h(\theta)) = \mathrm{sgn}(I(F(\theta)) - \theta).$$

That is, θ always moves towards $I(F(\theta))$. When the threshold is θ, $F(\theta)$ is the mass of players choosing strategy A, and so $I(F(\theta))$ is the bias of those players who are currently indifferent between strategies. If $\theta < I(F(\theta))$, players whose bias α lies between these values are currently playing strategy A (since $\alpha > \theta$) but would prefer to play strategy B (since $\alpha < I(F(\theta))$); the most dissatisfied players are those whose bias is exactly θ. The bias threshold therefore adjusts upward towards $I(F(\theta))$. Similarly, if $\theta > I(F(\theta))$, the threshold is driven downward. Only when $\theta = I(F(\theta))$ is society's behavior in equilibrium.

Figure 5 illustrates behavior adjustment in an equilibration game. Under equation (A), behavior adjustment leads to the unique equilibrium from any initial bias threshold. Moreover, if the preference distribution changes slightly, knocking the population slightly out of equilibrium, behavior adjustment quickly restores equilibrium play. This observation justifies our use of the equilibrium condition (E) in these games.

Figure 6 offers an example of behavior adjustment in a coordination game. Using this example and equation (A) as motivation, we call an equilibrium of a coordination game *stable* (*under behavior adjustment*) if F crosses I^{-1} from below at the equilibrium: that is, if

$$\mathrm{sgn}(\alpha - \theta) = \mathrm{sgn}(F(\alpha) - I^{-1}(\alpha))$$

[17] We can obtain similar results in a more complicated model without this restriction.

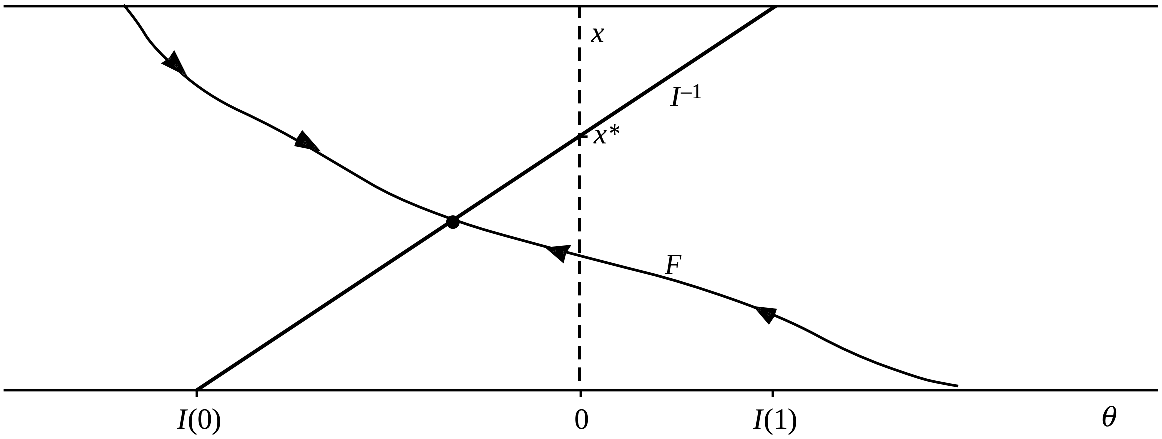

Figure 5: Behavior adjustment in an equilibration game.

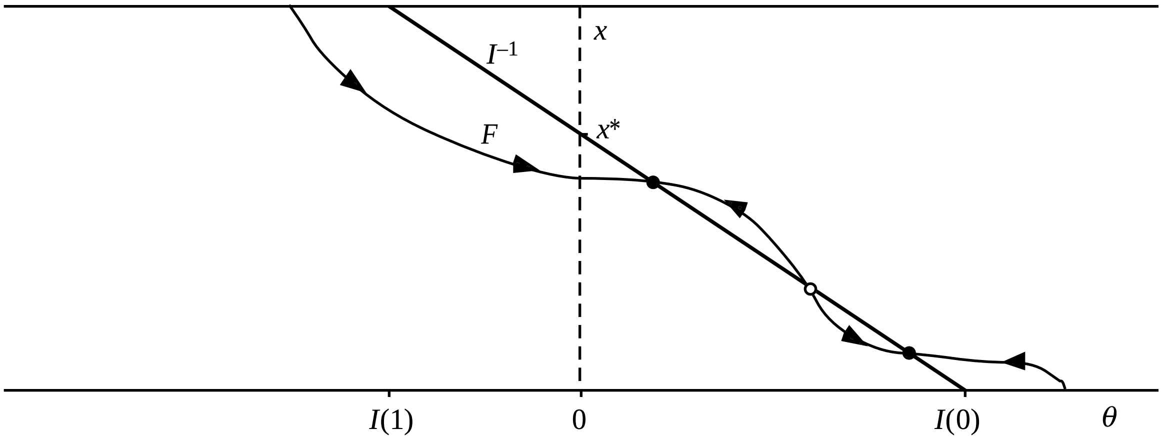

Figure 6: Behavior adjustment in a coordination game.

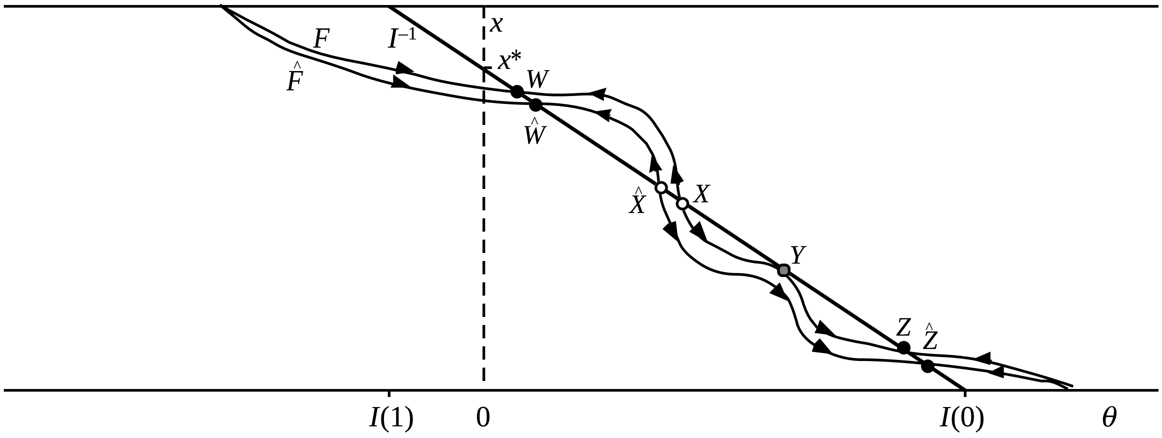

Figure 7: Stability of equilibria after a change in preferences.

for all α in a neighborhood of θ. We let $SE_G(f)$ denote the set of stable equilibria. Since $F'(\theta) = -f(\theta)$ and $(I^{-1})'(\theta) = -\Lambda(\theta)^{-1}$, a sufficient condition for an equilibrium θ to be stable under behavior adjustment is that $f(\theta) < \Lambda(\theta)^{-1}$: an equilibrium is stable if the number of indifferent players is smaller than the game's equilibrium alignment.

In Figure 7, we consider how aggregate behavior adjusts after a slight change in the preference distribution. Suppose that the initial preference distribution is F. In all of the equilibria pictured, fewer players choose strategy A than in the mixed equilibrium x^*. Hence, strategy B is more fit than strategy A, so evolution shifts preferences downward to a new distribution, $\hat{F}$. We first assume that this shift occurs without any change in the bias threshold, and then let the threshold adjust according to equation (A).

Suppose the population initially plays one of the stable equilibria, say $W = (\theta_W, x_W)$, and that preferences shift to $\hat{F}$. Then behavior adjustment from θ_W under $\hat{F}$ would lead the population to the new stable equilibrium $\hat{W} = (\theta_{\hat{W}}, x_{\hat{W}})$. In general, any stable equilibrium θ is robust to small changes in preferences, in the sense that a stable and locally unique equilibrium near θ will exist under the new preference distribution. Small changes in preferences can therefore be coupled with equally small behavior adjustments which maintain equilibrium play.

Now suppose that the population initially played the unstable equilibrium $X = (\theta_X, x_X)$. Then, after preferences shift, the bias threshold adjusts from θ_X to $\theta_{\hat{Z}}$. Because strategy adjustments lead away from them, unstable equilibria do not constitute reasonable predictions of play.

Finally, suppose that the population plays the tangent equilibrium $Y = (\theta_Y, x_Y)$. In this case, after preferences shift to $\hat{F}$ the equilibrium vanishes. As illustrated in the figure, behavior adjustment under equation (A) causes the bias threshold to increase until the stable equilibrium $\hat{Z}$ is reached.

In the next section, we show how preference evolution can convert a stable equilibrium into a tangent equilibrium, forcing behavior to jump abruptly to a new stable equilibrium. This new equilibrium is specified by the *jump function*, J;

$$J(f, \theta) = \begin{cases} \min\{\alpha > \theta : \ \alpha \in SE_G(f)\} & \text{if } \theta > 0; \\ \max\{\alpha < \theta : \ \alpha \in SE_G(f)\} & \text{if } \theta < 0. \end{cases}$$

Here, θ represents the tangent equilibrium which will be eliminated after preferences shift. If $\theta > 0$ (as is θ_Y), then behavior adjustment under equation (A) pushes the bias threshold to the right. The adjustment must not stop at another tangent equilibrium, as all such equilibria are eliminated when preferences shift. Therefore, the jump function picks the first *stable* equilibrium to the right of θ. Similarly, if the tangent equilibrium θ is to the left of 0, behavior adjustment will reduce the bias threshold; the jump function selects the first stable equilibrium to the left of θ.

It is not obvious that the jump function is well-defined. Establishing this is the key step in characterizing preference evolution in coordination games.

6. Coordination Games

6.1 The Extended Model

To model the evolution of preferences in coordination games we introduce two new assumptions.

(A4) $\theta_0 \in SE_G(f_0)$.

(A5) If $F_t(\theta) = I^{-1}(\theta)$, $F_t'(\theta) = (I^{-1})'(\theta)$, and $F_t''(\theta)$ exists, then $F_t''(\theta) \neq (I^{-1})''(\theta)$.

Assumption (A4) asks that the initial equilibrium be stable under behavior adjustment. Were coevolution to begin at an unstable equilibrium, changes in the population's composition caused by evolution would quickly disrupt it; subsequent behavior adjustment would quickly lead the population to a stable equilibrium. Assumption (A5) requires that if an equilibrium is reached at which the preference distribution and the indifference function have the same slope, they are strictly tangent at this point. Under this mild technical assumption, the inequality $f_t(\theta) < \Lambda(\theta)^{-1}$ is both sufficient and necessary for an equilibrium to be stable. More importantly, we establish in the Appendix (Proposition A15) that under assumption (A5), the jump function is well defined.[18]

The extended model consists of our original conditions on the evolutionary dynamics and equilibrium and a new condition concerning the continuity of the threshold adjustment process. To state this condition, we let $\theta_{t^-} = \lim_{s \uparrow t} \theta_s$ denote the left limit of a sequence of bias thresholds.

[18] It is possible to state assumption (A5) in terms of the model's initial conditions by using the function f^* defined below. We do not do so because this formulation is more difficult to interpret than is assumption (A5) as stated.

(D) For all $\alpha \in \mathbf{R}$, $\frac{d}{dt^+} f_t(\alpha) = f_t(\alpha)\, g_{\sigma(\alpha,\theta_t)}(x_t)$ and $f_t(\alpha)$ is continuous in t.

(E) $\theta_t \in E_G(f_t)$ for all t.

(C) θ_t is right continuous in t. If $\theta_{T^-} \in SE_G(f_T)$, then θ_t is also left continuous at time T. Otherwise, $\theta_T = J(f_T, \theta_{T^-})$.

Condition (C) formalizes the discussion of the previous section. As long as the equilibrium θ_t remains stable, it adjusts continuously in response to changes in the preference distribution. However, if a tangency is reached, equilibrium can not be maintained through continuous changes in θ_t, forcing a discrete behavior adjustment.

Theorem 2 characterizes preference evolution in coordination games. To state this result, we first define the functions f^* and K.

$$f^*(\theta) = \begin{cases} \dfrac{I^{-1}(\theta)\, f_0(\theta)}{F_0(\theta)} & \text{if } \theta \in (0, I(0)), \\ \dfrac{(1 - I^{-1}(\theta))\, f_0(\theta)}{(1 - F_0(\theta))} & \text{if } \theta \in (I(1), 0), \\ 0 & \text{otherwise.} \end{cases} \qquad K(\theta) = \begin{cases} (\theta, I(0)) & \text{if } \theta \in (0, I(0)), \\ (I(1), \theta) & \text{if } \theta \in (I(1), 0), \\ \varnothing & \text{otherwise.} \end{cases}$$

Theorem 2: *Suppose G is a coordination game, and let the initial condition* (f_0, θ_0) *satisfy assumptions* (A2) - (A5). *Then*:

(*i*: *Existence*) *There exists a unique solution trajectory*, $\{f_t, \theta_t\}_{t \geq 0}$, *to* (D), (E), *and* (C).

(*ii*: *Continuity*) θ_t *and* x_t *change continuously over time if and only if* $f^*(\theta) < \Lambda(\theta)^{-1}$ *for all* $\theta \in K(\theta_0)$.

(*iii*: *Limit Behavior*)

(*a*) *If* $x_0 < x^*$, *then* $\lim_{t\to\infty} x_t = 0$ *and* $\lim_{t\to\infty} \theta_t = I(0)$.

(*b*) *If* $x_0 = x^*$, *then* $x_t \equiv x^*$ *and* $\theta_t \equiv 0$.

(*c*) *If* $x_0 > x^*$, *then* $\lim_{t\to\infty} x_t = 1$ *and* $\lim_{t\to\infty} \theta_t = I(1)$.

Part (*i*) establishes the existence and uniqueness of the solution trajectories. The existence of the continuous portions of the trajectories is established using the ODE reduction discussed in Section 4. To establish the existence of the entire trajectory, we must also show that the jump function J is well defined. In the Appendix

(Proposition A15), we offer an inductive argument which considers the equilibria which lie beyond the launching point of a jump one by one. The argument shows that under assumption (A5), a jump need only pass over a finite number of tangent equilibria before reaching the first stable equilibrium.

Part (*ii*) presents a necessary and sufficient condition for jumps to occur; we discuss this result in Section 6.2. Part (*iii*) characterizes limit behavior. If $x_0 = x^*$, then the coevolutionary trajectory is degenerate, as the population forever remains at the mixed equilibrium of the fitness game.[19] Otherwise, the population eventually coordinates on either strategy *A* or strategy *B* depending on the initial conditions.

We sketch an example of preference evolution in a coordination game in Figure 8. In this example, the number of players who initially play strategy *A* is smaller than the number who play *A* in the mixed equilibrium. Since *G* is a coordination game, it follows that strategy *B* has the higher fitness. Consequently, biases below θ_0 grow more prevalent at the expense of the others. The primary effect of this change in the bias distribution is to increase the number of players choosing strategy *B*. This in turn spurs a secondary effect. Because *G* is positively aligned, that more players choose strategy *B* makes strategy *B* even more attractive. Hence, players who had been indifferent between strategies switch to strategy *B*. The primary and secondary effects of preference evolution reinforce one another, driving the bias threshold rightward.

The derivation from Section 4 shows that at times that strategy adjustment is continuous, aggregate behavior adjusts according to the equation

$$\text{(B)} \qquad \dot{x}_t = \frac{x_t \, g_A(x_t)}{1 - \Lambda(\theta_t) f_t(\theta_t)}.$$

This is the same law of motion which governed strategy adjustment in equilibration games. Since $f_0(\theta_0) < \Lambda(\theta_0)^{-1}$ by our stability assumption, the denominator of this expression is positive and less than one; hence, strategy adjustment is faster than in the standard evolutionary model defined by $\dot{x}_t = x_t \, g_A(x_t)$. The higher the density at the bias threshold (subject to $f_t(\theta_t) < \Lambda(\theta_t)^{-1}$), the more players who switch from *A* to *B* at time *t*, and hence the greater the acceleration of the strategy adjustment process. Higher alignments also yield faster strategy adjustment.

[19] By assumption (A4), this equilibrium must be stable under behavior adjustment. For an example, see Figure 10.

Evolution proceeds smoothly as long as $f_t(\theta_t)$ stays below $\Lambda(\theta_t)^{-1}$, so that θ_t remains a stable equilibrium. In our example, $f_t(\theta_t)$ approaches $\Lambda(\theta_t)^{-1}$ as t approaches T. Hence, the speed of adjustment, $|\dot{x}_t|$, grows without bound. At the limit, $f_T(\theta_{T^-}) = \Lambda(\theta_{T^-})^{-1}$: the preference distribution becomes tangent to the indifference line. Because the limit equilibrium is unstable, a round of behavior adjustment ensues, sending the population to θ_T, the first stable equilibrium to the right of θ_{T^-}. From this stable equilibrium continuous evolution can resume.

We label the limit bias distribution F_∞, and the limit bias threshold $\theta_\infty = I(0)$. The limit threshold dictates that strategy B is chosen by all players for whom it is not dominated. At the limit distribution, the mass of such players is one: ultimately, all players whose bias towards A is so large that they can never be reconciled to play B vanish from the population.

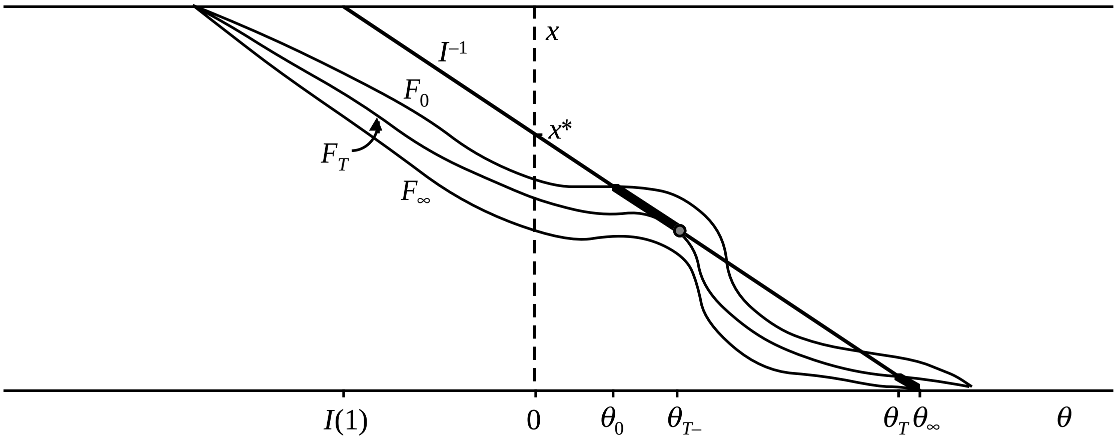

Figure 8: Evolution of preferences in a coordination game.

The reason that jumps can only occur in coordination games can be understood by considering the equilibrium correspondence, $E_G(\cdot)$.[20] In equilibration games, the equilibrium correspondence is a continuous function: as we continuously alter the preference distribution, the unique equilibrium of the game also varies continuously. This guarantees that in such games, a continuous trajectory for the bias threshold can always be found. In contrast, when G is a coordination game, $E_G(\cdot)$ is an upper hemicontinuous correspondence. Since the correspondence fails to be lower hemicontinuous, there are equilibria which can be disrupted by arbitrarily small changes in the preference distribution. These are precisely the equilibria at which the preference distribution and the indifference line are tangent. Our

[20] A formal analysis can be found in the Appendix (Proposition A1).

analysis shows that the preference trajectory often passes through discontinuities in the equilibrium correspondence, forcing instantaneous jumps to new equilibria.[21]

Our analysis is based on the assumption that behavior adjusts to equilibrium an order of magnitude faster than preferences evolve. However, our results are not overly dependent on this assumption. We show in the Appendix that the present model is the limiting case of a model in which the relative rates of adjustment are large but finite, and in which disequilibrium behavior is possible.

6.2 Which Initial Conditions Lead to Rapid Social Change?

Theorem 2 (*ii*) presents a necessary and sufficient condition for continuous strategy adjustment which is stated in terms of the function f^*. We show in the Appendix (Lemma A10) that $f^*(\theta_t)$ is equal to the preference density at bias θ_t and time t: $f^*(\theta_t) = f_t(\theta_t)$. In other words, $f^*(\theta)$ is the preference density at the equilibrium when the equilibrium is θ. If $f^*(\theta) < \Lambda(\theta)^{-1}$ for all $\theta \in (\theta_0, I(0))$, then the slope of the preference distribution at θ_t is always less than the slope of the indifference line, and so it is always possible for evolution to proceed smoothly. On the other hand, if $\Lambda(\theta_t)f^*(\theta_t)$ converges to one during the course of evolution, the preference distribution and the indifference line become tangent, forcing a jump.

We can also offer separate necessary and sufficient conditions for jumps solely in terms of the initial bias density f_0.

Corollary 3: *If* $f_0(\theta) < \min\limits_{\theta\in[I(1),I(0)]} \Lambda(\theta)^{-1}$ *for all* θ, *strategy adjustment is continuous. On the other hand, if* $E_G(f_0) \cap K(\theta_0) \neq \emptyset$, *strategy adjustment is discontinuous.*

The first claim of the corollary shows that if the preference distribution is sufficiently diffuse relative to the game's alignment, no jumps will occur. Under this condition, the preference density at the equilibrium always remains small enough that the speed of evolution stays finite.

[21] We should note that if play begins at an unstable equilibrium (i.e., if $f_0(\theta_0) > \Lambda(\theta_0)^{-1}$), equation (B) implies that behavior will adjust in the direction opposite to that in the standard evolutionary model. For example, if $x_0 > x^*$, then despite the fact that strategy B has the higher expected fitness, the mass of players playing strategy B will *decrease* over time. However, equilibria which are not stable under strategy adjustment should not be expected to persist in the face of changes in the population's composition. We therefore regard this point as a technical curiosity.

The second claim states a sufficient condition for jumps in terms of the equilibria which exist at the initial moment of play. If at the onset of play there are equilibria further from zero than θ_0, a jump is bound to occur.

When will such equilibria exist? Consider Figure 8, in which $\theta_0 > 0$. Since this initial equilibrium is stable ($f_0(\theta_0) < \Lambda(\theta_0)^{-1}$), the preference distribution crosses the indifference line from below. Therefore, for another equilibrium to lie to the right of θ_0, there must be an interval of biases between θ_0 and $I(0)$ over which the initial preference distribution is relatively dense. That is, there must be a cluster of players whose bias at first leads them to strategy A, but whose bias towards A is moderate enough that they are ultimately willing to play strategy B. Our analysis shows that the existence of such a cluster is enough to guarantee a discrete change in aggregate behavior.

7. Purification and Evolution

In an n-player normal form game, whenever a player's equilibrium strategy is mixed he is indifferent between this strategy and all others with the same support. This raises the question of why we should expect him to randomize, and moreover to randomize in the precise fashion that his equilibrium strategy dictates. In his 1973 paper, Harsanyi offers a resolution to this puzzle. He does so by augmenting the original normal form game by adding to each player's payoffs a small, idiosyncratic noise term which depends on the realized action profile. While the distributions of the payoff noises are commonly known, they are independent, and players only learn the realizations of their own noise terms. Each pure Bayesian strategy in the perturbed game can be identified with a mixed strategy in the original game, with each player's noise terms serving as his randomizing device. Harsanyi proves that all equilibria of generic normal form games are the limits of sequences of equilibria of perturbed games whose noise terms become arbitrarily small. Importantly, in the equilibria of the perturbed games, almost all noise realizations induce a strict preference for a single action. In other words, small payoff noises eliminate the indifference which makes mixed equilibria unstable.

There are close formal connections between Harsanyi's model and our own. Restrict attention to the models' common ground: 2 x 2 symmetric normal form games,[22] and payoff disturbances which only depend on players' own actions. Although Harsanyi studies a game of incomplete information while we study a population game, it is easy to see that if we identify payoff noises with biases, the two models are formally identical. Hence, any Bayesian equilibrium under noise distribution F can be interpreted as a population equilibrium under the fixed bias distribution F.

Suppose further that $F(0) = x^*$, and consider the equilibrium $X = (0, x^*)$. If we interpret X as a purified equilibrium under noise distribution F, X is an equilibrium of the perturbed game which perfectly mirrors the equilibrium of the underlying game. On the other hand, if we interpret X as a population equilibrium under bias distribution F, then since $\phi_A(x^*) = \phi_B(x^*)$, X is an equilibrium at which the distribution of preferences is at rest. Hence, noise distributions which yield exact purification correspond to bias distributions which face no selection pressure.

Harsanyi’s purification results offer a basis for belief in nearly any mixed equilibrium. The discussion above suggests the following question: if the underlying payoffs are interpreted as evolutionary fitnesses, can purified mixed equilibria always be given an evolutionary justification? In other words, given any mixed equilibrium x^* of the fitness game G, can we always find a distribution F under which x^* is an evolutionarily stable outcome?

While if attention is restricted to equilibration games the answer to this question is yes, for coordination games the answer is no. While this in itself may seem unsurprising, the reason behind this answer is perhaps more subtle than it first appears to be. As before, we consider the mixed equilibrium x^* of a 2 x 2 symmetric coordination game G, and let F be a preference distribution satisfying $F(0) = x^*$, so that $X = (0, x^*)$ is an equilibrium under this preference distribution. Since $\phi_A(x^*) = \phi_B(x^*)$, there is no evolutionary pressure on F to change.

Nevertheless, we should not expect this equilibrium to persist. If F crosses I^{-1} from above at the equilibrium, as pictured in Figure 9, then the equilibrium is unstable under behavior adjustment: a slight change in behavior will lead the population away from the equilibrium.

[22] Harsanyi does not explicitly consider games with a single player role, but his results can easily be extended to this case.

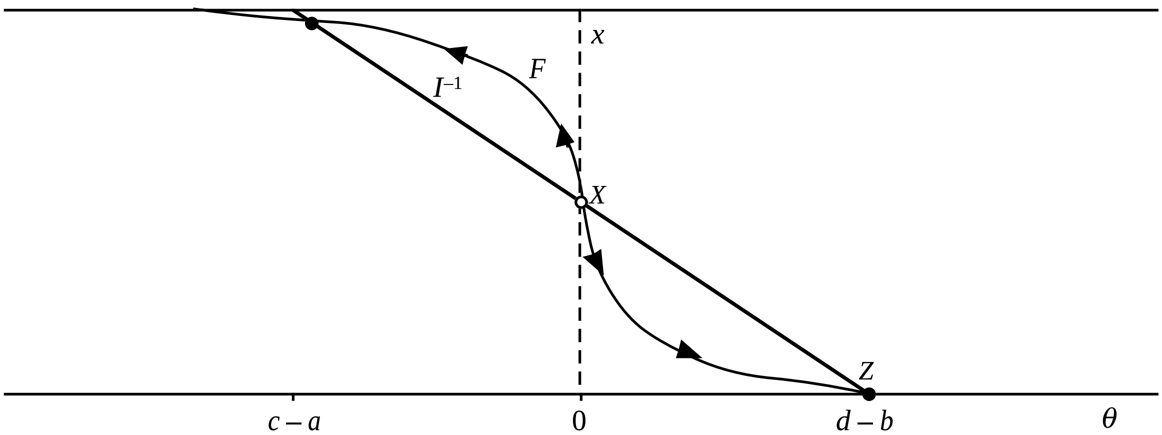

Figure 9: An equilibrium which is not robust to behavior adjustments or preference shifts.

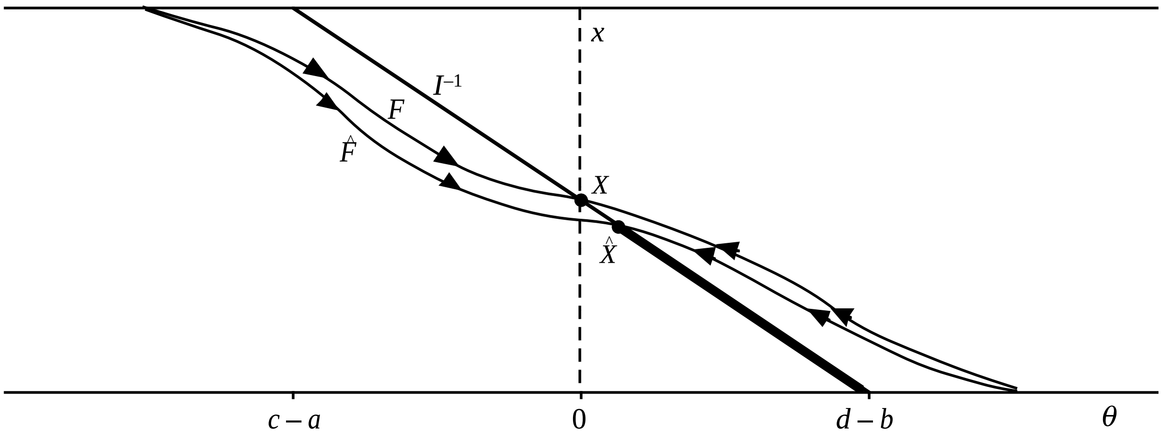

Figure 10: An equilibrium which is robust to behavior adjustments but not to preference shifts.

In contrast, if F crosses I^{-1} from below, as illustrated in Figure 10, the equilibrium is stable under behavior adjustment: the population returns to the equilibrium after any slight change in behavior. But suppose there is a slight change in society's preferences to $\hat{F}$, so that the unique equilibrium under the new preferences, $\hat{X}$ = $(\theta_{\hat{X}}, x_{\hat{X}})$, satisfies $x_{\hat{X}} \neq x^*$. At this equilibrium, one of the two strategies must have a strict fitness advantage. Hence, Theorem 2 implies that preference evolution will lead the population to coordinate on a single strategy. Therefore, while the equilibrium $X = (0, x^*)$ is stable if preferences are held fixed at F, a small change in the preference distribution will lead the population away from the equilibrium.[23]

[23] A preference shift will also disrupt the equilibrium in Figure 9. What is important is that *only* a preference shift can disrupt the equilibrium in Figure 10; a change in behavior will not suffice.

In summary, our analysis of coordination games shows that if payoff disturbances are large, mixed equilibrium *behavior* can be stable, but that the disturbance distribution itself cannot be stable.

8. Conclusion

We analyze an explicitly dynamic model of preference evolution, establishing the existence and uniqueness of the evolutionary solution paths. To keep the analysis as simple as possible, we have restricted attention to games with two strategies and to preferences which can be represented in terms of biases. There are many applications of large population games in which players face a binary choices,[24] and biases seem an especially natural form of variation in individual preferences. Nevertheless, understanding the dynamics of preference evolution in more general strategic settings and under broader classes of preferences is an important topic for future research.

Appendix

A.1 Speed Limits and Disequilibrium Strategy Adjustment

Our primary model requires that behavior adjust arbitrarily more quickly than preferences evolve. While it is natural to assume that the former process is faster than the latter, assuming that the relative rate of adjustment is infinite may seem

It is worth noting that pure equilibria which are stable under behavior adjustment are robust to small changes in the preference distribution. Consider the equilibrium $Z = (d - b, 0)$ in Figure 9, at which all players choose strategy B. This equilibrium is stable under behavior adjustment, and any sufficiently small change in the preference distribution will yield a new stable equilibrium near Z. If some players play strategy A at this new equilibrium, preference evolution will select against these players' biases until all players again choose strategy B.

[24] Many authors (e.g., Kandori, Mailath, and Rob (1993)) have used such games to study consumer technology choice. Durlauf (1997) mentions out-of-wedlock births, school attendance, and criminal activity as examples of socioeconomic issues which can be studied using large population binary choice models. Kuran's (1989, 1991, 1995) model of political and cultural revolutions is also of this form.

too extreme. We now demonstrate that weakening this assumption has only a minor impact on our analysis.

Specifically, we impose a speed limit on the strategy adjustment process. That is, we assume that the speed of strategy adjustment, $|\dot{x}_t|$, has some fixed upper bound M. When M is large, behavior may change much more quickly than preferences, but always finitely so.

Define the bounding function $B^M : \mathbf{R} \to \mathbf{R}$ by

$$B^M(x) = \begin{cases} M & \text{if } x \in (M, \infty), \\ x & \text{if } x \in [-M, M], \\ -M & \text{if } x \in (-\infty, M). \end{cases}$$

We formalize the model of coevolution with speed limits as follows.

(D) For all $\alpha \in \mathbf{R}$, $\frac{d}{dt^+} f_t(\alpha) = f_t(\alpha)\, g_{\sigma(\alpha,\theta_t)}(x_t)$ and $f_t(\alpha)$ is continuous in t.

(SL1) If $\theta_t \in E_G(f_t)$, then $\frac{d}{dt^+} x_t = B^M\left(\dfrac{x_t\, g_A(x_t)}{1 - \Lambda(\theta_t) f_t(\theta_t)}\right)$.

(SL2) If $\theta_t \notin E_G(f_t)$, then $\frac{d}{dt^+} x_t = M \operatorname{sgn}(\theta_t - I(F_t(\theta_t)))$.

(SL3) The trajectory $x_{(\cdot)}$ is continuous.

Preference evolution, modeled by condition (D), is unchanged. To interpret condition (SL1), recall from equation (B) that $x_t\, g_A(x_t) / (1 - \Lambda(\theta_t) f_t(\theta_t))$ is the rate of strategy adjustment which maintains equilibrium in our original model. When behavior is in equilibrium, condition (SL1) states that that behavior adjusts to preserve equilibrium whenever doing so does not require breaking the speed limit. If a point is reached at which maintaining equilibrium would require a speed limit violation, equilibrium play is abandoned. While behavior is out of equilibrium, condition (SL2) asks that strategy adjustment occur at rate M until equilibrium is restored.

Since a formal study of the speed limit model requires only a minor modification of our prior analysis, we do not provide one here.[25] Instead, in Figure 11 we revisit the initial conditions from Figure 8 and consider coevolution with speed limits imposed. Until we near the time at which the jump would have

[25] It is worth noting, however, that because assumption (A5) only imposes restrictions at times that jumps occur, in the speed limit model it is unnecessary. Also, since in equilibration games coevolution only slows strategy adjustment down, adding sufficiently high speed limits has no effect on the model.

occurred, coevolution proceeds just as before. However, at some time u before the jump can occur, the speed limit M is reached. A period of disequilibrium follows, during which players switch to strategy B at rate M. During this time, the preference distribution continues to evolve as condition (D) dictates. Eventually, at some time v, behavior returns to equilibrium, and coevolution proceeds as before.[26]

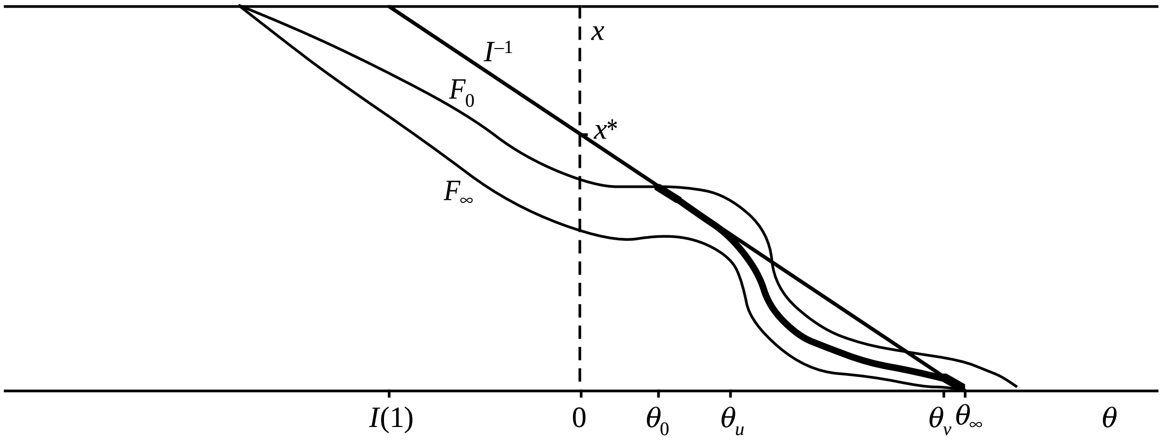

Figure 11: Evolution of preferences under a speed limit.

What happens as we increase the speed limit? The key observation needed to answer this question is that the disequilibrium portion of trajectories resulting from different speed limits are nested: in Figure 11, increasing M leads to a new disequilibrium trajectory which is strictly above the old one on (θ_u, θ_v). It follows that as we increase the speed limit, the bias at which equilibrium is abandoned, θ_u, moves to the right; the bias at which equilibrium play resumes, θ_v, moves to the left; and the time spent in disequilibrium, $v - u$, decreases. Let θ_{T^-} and θ_T be the endpoints of the jumps in the original model. Applying the observation noted above, it is not difficult to show that as M approaches infinity, u approaches T, θ_u approaches θ_{T^-}, θ_v approaches θ_T, and $v - u$ approaches zero. Therefore, we can view our original model as the limiting case of the speed limit model as the speed limit becomes arbitrarily large.

[26] When equilibrium is restored at time v, all biases above θ_v have always induced play of strategy A, and so have always shrunk at the same rate. Therefore, although the time at which bias θ_v is reached will be later than in the original model, the speed of strategy adjustment thereafter is the same.

A.2 Proofs

We begin by studying the continuity of the equilibrium correspondence $E_G(\cdot)$. Define the distance between two preference distributions to be the maximal distance between their corresponding decumulative distribution functions:

$$|f_n - f_m| = \sup_{\lambda} |F_n(\lambda) - F_m(\lambda)|.$$

Proposition A1: (*i*) *If G is a coordination game, then $E_G(\cdot)$ is non-empty and upper hemicontinuous.* (*ii*) *If G is an equilibration game, then $E_G(\cdot)$ is a continuous function.*

Proof: We begin with a lemma:

Lemma A2: *If $f_n \to f$, $\theta_n \to \theta$, and $\theta_n \in E_G(f_n)$ for all n, then $\theta \in E_G(f)$.*

Proof: Since $\theta_n = I(F(\theta_n))$ for all n, we see that

$$|\theta - I(F(\theta))| \leq |\theta - \theta_n| + |I(F_n(\theta_n)) - I(F(\theta_n))| + |I(F(\theta_n)) - I(F(\theta))|$$

Since all of the terms on the right hand side approach zero as n approaches infinity, $\theta = I(F(\theta))$. ❑

As $E_G(\cdot)$ is clearly non-empty, part (*i*) of the proposition follows directly from the lemma.

To prove part (*ii*), let G be negatively aligned, so that $E_G(\cdot)$ is single valued. Suppose that $f_n \to f$ and that $\theta_n \in E_G(f_n)$ for all n, and let θ^* be the lone element of $E_G(f)$. We want to show that $\theta_n \to \theta^*$. Were this not true, we could find an $\varepsilon > 0$ and a subsequence θ_{n_k} such that $\left|\theta_{n_k} - \theta^*\right| > \varepsilon$ for all k. Since all θ_{n_k} lie in the compact set $[I(0), I(1)]$, this subsequence has in turn a convergent subsequence. Lemma A2 implies that its limit is θ^*, contradicting the definition of the subsequence. ■

The next result shows that all continuous portions of solution trajectories to equations (D) and (E) can be expressed in terms of solutions to a two dimensional ODE. The two variables in the ODE are the bias threshold θ_t and the aggregate percentage growth rate γ_t. The latter variable measures the aggregate growth of those biases which have always prompted play of the strategy which the marginal players are abandoning. For example, on solution trajectories along which strategy *B* is more fit than strategy *A*, γ_t measures the percentage change in the representation of any bias which has always induced play of strategy *A*. This implies that the density of any bias α that has always played strategy *A* is given by $f_t(\alpha) = f_0(\alpha)\ \gamma_t$. If we can keep track of which biases play strategy *A* using the threshold θ_t, then γ_t contains enough information about changes in the bias distribution to allow us to perform the reduction.

The direction (*ii*) ⇒ (*i*) shows that solutions to an ODE in the variables θ_t and γ_t form the basis for solutions to (D) and (E). This is the key to the proof of the existence of solutions to the latter equations. To prove that these solutions are unique, we use the implication (*i*) ⇒ (*ii*), which shows that any such solution also solves the ODE. The proofs begin by showing that under either set of equations, the bias threshold θ_t must change monotonically over time. Once this is established, algebra and calculus are used to move between the two sets of equations.

Theorem A3 (*ODE Reduction*): *Fix a game G and an initial condition* (f_0, θ_0) *which satisfies assumptions* (A1), (A2), *and, if G is a coordination game,* (A4). *Then the following two statements are equivalent*:

(*i*) *The continuous trajectory* $\{f_t, \theta_t\}_{t\in[0,T)}$ *satisfies* $f_t(\theta_t) \neq \Lambda(\theta_t)^{-1}$ *for all t and solves*

(D) For each $\alpha \in \mathbf{R}$, $\frac{d}{dt^+} f_t(\alpha) = g_{\sigma(\alpha,\theta_t)}(x_t)\, f_t(\alpha)$ and $f_{(\cdot)}(\alpha)$ is continuous;

(E) $\theta_t = E_G(f_t)$ for all *t*.

Furthermore, the trajectory $\{\gamma_t\}_{t\in[0,T)}$ *is given by*

(G)
$$\gamma_t = \begin{cases} \exp\left(\int_0^t g_A(x_s)ds\right) & \text{if } \Lambda(\theta_0)\theta_0 \geq 0, \\ \exp\left(\int_0^t g_B(x_s)ds\right) & \text{otherwise.} \end{cases}$$

(*ii*) *The* C^1 *trajectory* $\{\theta_t, \gamma_t\}_{t\in[0,T)}$, *which satisfies* $f_0(\theta_t)\gamma_t \neq \Lambda(\theta_t)^{-1}$ *for all t and* $\gamma_0 = 1$, *solves*

$$\text{(C1)}\qquad \dot{\theta}_t = \begin{cases} \dfrac{F_0(\theta_t)\,\gamma_t\; g_A\big(I^{-1}(\theta_t)\big)}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}} & \text{if } \Lambda(\theta_0)\theta_0 \geq 0, \\ \dfrac{\big(1-F_0(\theta_t)\big)\,\gamma_t\; g_B\big(I^{-1}(\theta_t)\big)}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}} & \text{otherwise;} \end{cases}$$

$$\text{(C2)}\qquad \dot{\gamma}_t = \begin{cases} g_A\big(I^{-1}(\theta_t)\big)\,\gamma_t & \text{if } \Lambda(\theta_0)\theta_0 \geq 0, \\ g_B\big(I^{-1}(\theta_t)\big)\,\gamma_t & \text{otherwise;} \end{cases}$$

and the trajectory $\{f_t\}_{t\in[0,T)}$ *is given by*

$$\text{(C3)}\qquad f_t(\alpha) = f_0(\alpha)\exp\Big(\int_0^t g_{\sigma(\alpha,\theta_s)}(I^{-1}(\theta_s))ds\Big).$$

Proof of (*ii*) ⇒ (*i*): We state the proof for the case in which $\Lambda(\theta_0)\theta_0 \geq 0$; the proof of the other case is similar.

Since the trajectory $\{\theta_t, \gamma_t\}_{t\in[0,T)}$ solves (C1) and (C2), this trajectory is continuous. We now establish that θ_t is increasing on $[0, T)$. First, note that γ_t is strictly positive for finite t. Therefore, since $f_0(\theta_t)$, γ_t, and $\Lambda(\theta_t)^{-1}$ are continuous in t, the denominator of the right hand side of (C1) cannot switch signs without passing through zero, which would render (C1) undefined. Hence, the right hand side of (C1) can only switch signs if its numerator switches signs, which can only occur if $g_A(I^{-1}(\theta_t))$ switches signs.

A simple calculation shows that when $\Lambda(\theta_0)\theta_0 \geq 0$, $g_A(I^{-1}(\theta_0)) \geq 0$. Since θ_t is continuous on $[0, T)$, $g_A(I^{-1}(\theta_t))$ can only become negative if it passes through zero. But if $g_A(I^{-1}(\theta_t)) = 0$ at some time t, it equals zero at all future times, and θ_t is constant over these times. Hence, $g_A(I^{-1}(\theta_t)) \geq 0$ for all $t \in [0, T)$, and so θ_t is increasing on $[0, T)$.

Because θ_t is increasing on $[0, T)$, a player whose bias at time t is at least θ_t has always played strategy A; that is, $\sigma(\alpha, \theta_s) = A$ for all $\alpha \geq \theta_t$ and $s \leq t$. Substituting this expression into equation (C3), we see that

$$f_t(\alpha) = f_0(\alpha)\exp\Big(\int_0^t g_A(I^{-1}(\theta_s))ds\Big) \qquad \text{for all } \alpha \geq \theta_t. \tag{1}$$

Since equation (C2) and the initial condition $\gamma_0 = 1$ imply that

$$\gamma_t = \exp\left(\int_0^t g_A(I^{-1}(\theta_s))ds\right),$$

we see that

$$f_t(\alpha) = f_0(\alpha)\,\gamma_t \text{ for all } \alpha \geq \theta_t. \tag{2}$$

Hence, the assumption in statement (*ii*) that $f_0(\theta_t)\gamma_t \neq \Lambda(\theta_t)^{-1}$ implies that $f_t(\theta_t) \neq \Lambda(\theta_t)^{-1}$, as required by statement (*i*).

Moreover, differentiating equation (1) yields

$$\dot{f}_t(\alpha) = f_t(\alpha)\,g_A(I^{-1}(\theta_t))\text{for all } \alpha > \theta_t. \tag{3}$$

Thus, rewriting equation (C1) and then substituting equations (2) and (3) yields

$$\begin{aligned}
\dot{\theta}_t &= \frac{F_0(\theta_t)\,\gamma_t\;g_A(I^{-1}(\theta_t))}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}} \\
&= \frac{\int_{\theta_t}^{\infty} f_0(\alpha)\,\gamma_t\;g_A(I^{-1}(\theta_t))d\alpha}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}} \\
&= \frac{\int_{\theta_t}^{\infty} f_t(\alpha)\,g_A(I^{-1}(\theta_t))d\alpha}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}} \\
&= \frac{\int_{\theta_t}^{\infty} \dot{f}_t(\alpha)d\alpha}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}}.
\end{aligned}$$

We rearrange this equation to find that

$$-\Lambda(\theta_t)^{-1}\,\dot{\theta}_t = \int_{\theta_t}^{\infty} \dot{f}_t(\alpha)d\alpha - \dot{\theta}_t f_t(\theta_t).$$

Observe that $\frac{d}{d\theta}I^{-1}(\theta) = -\Lambda(\theta)^{-1}$. Therefore, by integrating over t we obtain

$$I^{-1}(\theta_t) = \int_{\theta_t}^{\infty} f_t(\alpha)d\alpha \;+ C,$$

where C is a constant. Substituting $x_t = \int_{\theta_t}^{\infty} f_t(\alpha)d\alpha$ and composing the result with I yields

$$\theta_t = I(x_t + C).$$

Since $\theta_0 \in E_G(f_0)$, $\theta_0 = I(x_0)$. Therefore, since I is one-to-one, $C = 0$, and we conclude that

$$\theta_t = I(x_t) = I(F_t(\theta_t)).$$

This is condition (E).

To complete this part of the proof, observe that condition (E) implies that $x_t = I^{-1}(\theta_t)$. Substituting this expression into equations (C2) and (C3) yields

$$\begin{aligned} \dot{\gamma}_t &= g_A(x_t)\,\gamma_t \quad \text{and} \\ f_t(\alpha) &= f_0(\alpha)\exp\left(\int_0^t g_{\sigma(\alpha,\theta_s)}(x_s)ds\right). \end{aligned}$$

Integrating the former expression and differentiating the latter with respect to time yield equations (G) and (D). ❑

Proof of (*i*) ⇒ (*ii*): We divide the proof into two cases.

Case 1: *G is an equilibration game*. We prove this case under the assumption that $\theta_0 \leq 0$; if $\theta_0 > 0$ the proof is similar. We begin with three lemmas.

Lemma A4: $g_B(x_t) \leq 0$ *for all* $t \in [0, T)$.

Proof: Observe that $g_B(x_t) < 0$ whenever $x_t \in (0, x^*)$, and that $\theta_0 < 0$ implies that $x_0 \in (0, x^*)$. Moreover, θ_t is continuous on $[0, T)$ by assumption; since $x_t \equiv I^{-1}(\theta_t)$ by condition (E), x_t is continuous on $[0, T)$, too. Therefore, for x_t to leave $(0, x^*)$, there must be a time τ at which $x_\tau \in \{0, x^*\}$. But if this occurs, $g_B(x_\tau) = 0$, and so condition (D) implies that $g_B(x_t) = 0$ for all $t \in [\tau, T)$. Therefore, $g_B(x_t)$ can never become strictly positive. ❑

Lemma A5: θ_t *is increasing on* $[0, T)$.

Proof: Fix two times $u, v \in [0, T)$ with $u < v$. We establish that $\theta_u \leq \theta_v$ by showing that $u \in \underset{t\in[u,v]}{\arg\min}\, \theta_t$. Suppose to the contrary that $u \notin \underset{t\in[u,v]}{\arg\min}\, \theta_t$. Then if $\tau \in \underset{t\in[u,v]}{\arg\min}\, \theta_t$, then $\theta_\tau < \theta_u$, and moreover, $\sigma(\alpha, \theta_t) = B$ for all $\alpha < \theta_\tau$ and $t \in [u, v]$.

Hence, equation (D) and Lemma A4 imply that

$$f_\tau(\alpha) \;=\; f_u(\alpha)\exp\Bigl(\int_u^\tau g_B(x_s)ds\Bigr) \le f_u(\alpha) \quad \text{for all } \alpha < \theta_\tau. \tag{4}$$

Therefore,

$$\begin{aligned}
F_\tau(\theta_\tau) &= 1 - \int_{-\infty}^{\theta_\tau} f_\tau(\alpha)d\alpha && \text{by definition}\\
&\ge 1 - \int_{-\infty}^{\theta_\tau} f_u(\alpha)d\alpha && \text{by equation (4)}\\
&= F_u(\theta_\tau) && \text{by definition}\\
&\ge F_u(\theta_u) && \text{since } F_u \text{ is decreasing}\\
&= I^{-1}(\theta_u) && \text{since } \theta_u \in E_G(f_u)\\
&> I^{-1}(\theta_\tau) && \text{since } I^{-1} \text{ is strictly increasing.}
\end{aligned}$$

Thus, $F_\tau(\theta_\tau) > I^{-1}(\theta_\tau)$, contradicting condition (E). ❑

Lemma A6: *If θ_t is increasing on* $[0, T)$, *then for all* $\alpha \ge \theta_t$,

$$f_t(\alpha) = f_0(\alpha)\gamma_t \qquad \text{and} \qquad F_t(\alpha) = F_0(\alpha)\gamma_t. \tag{5) and (6}$$

Proof: Since θ_t is increasing in t, $\sigma(\alpha, \theta_s) = A$ for all $\alpha \ge \theta_t$ and $s \le t$. Therefore, for each $\alpha \ge \theta_t$, integrating equation (D) and substituting condition (G) yields

$$\begin{aligned}
f_t(\alpha) &= f_0(\alpha)\exp\Bigl(\int_0^t g_{\sigma(\alpha,\theta_s)}(x_s)ds\Bigr)\\
&= f_0(\alpha)\exp\Bigl(\int_0^t g_A(x_s)ds\Bigr)\\
&= f_0(\alpha)\,\gamma_t.
\end{aligned}$$

This is equation (5). Integrating over $\lambda \ge \alpha$ yields equation (6). ❑

We now complete the proof of Case 1. Since $f_t(\theta_t) \ne \Lambda(\theta_t)^{-1}$ by assumption, equation (5) implies that $f_0(\alpha)\,\gamma_t \ne \Lambda(\theta_t)^{-1}$ as required by statement (*ii*). Equilibrium condition (E) implies that $\theta_t = I(F_t(\theta_t))$. Using equation (6), we rewrite this as

$$I^{-1}(\theta_t) = \gamma_t F_0(\theta_t)$$

Both sides of this equation are continuously differentiable in θ_t, and γ_t is a C^1 function of t on [0, T). Therefore, the Implicit Function Theorem implies that θ_t is C^1 on [0, T) and that we can determine its derivative through implicit differentiation. Recalling that F_0 is a decumulative distribution, we differentiate the previous expression to obtain

$$-\Lambda(\theta_t)^{-1}\,\dot{\theta}_t = \dot{\gamma}_t\,F_0(\theta_t) - \gamma_t\,\dot{\theta}_t\,f_0(\theta_t).$$

Differentiating equation (G) yields

$$\dot{\gamma}_t = g_A(x_t)\,\gamma_t.$$

Combining these last two expressions and rearranging, we see that

$$\dot{\theta}_t = \frac{F_0(\theta_t)\,\gamma_t\,g_A(x_t)}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}}.$$

Since $x_t = I^{-1}(\theta_t)$ by condition (E), we conclude that equation (C1) holds. Finally, equations (C2) and (C3) follow directly from conditions (D), (E), and (G). ❑

Case 2: *G is a coordination game.* We prove this case under the assumption that $\theta_0 \geq 0$; if $\theta_0 < 0$ the proof is similar.

Lemma A7: $g_B(x_t) \geq 0$ *for all* $t \in$ [0, T).

Proof: Analogous to that of Lemma A4.

Lemma A8: θ_t *is increasing on* [0, t).

Proof: Suppose to the contrary that θ_t is not increasing on [0, T). Let $\overline{\theta}_t = \max_{s\in[0,t]} \theta_t$; our supposition implies that $\tau = \inf\{t\text{: } \overline{\theta}_t \neq \theta_t\} < T$. Since θ_t is increasing through time τ, $\sigma(\alpha, \theta_t) = A$ for all $\alpha \geq \theta_t$ and $t \leq \tau$. Consequently, integrating equation (D) yields

$$f_t(\theta_t) = f_0(\theta_t)\exp\left(\int_0^t g_{\sigma(\theta_t,\theta_s)}(x_s)ds\right)$$

$$= f_0(\theta_t)\exp\left(\int_0^t g_A(x_s)ds\right)$$

for all $t \leq \tau$. Since θ_t is continuous in t by assumption, $f_t(\theta_t)$ is continuous in t over the interval $[0, \tau]$.

Assumption (A4) and the condition in statement (*i*) that $f_t(\theta_t) \neq \Lambda(\theta_t)^{-1}$ for all $t \in [0, T)$ imply that $f_0(\theta_0) < \Lambda(\theta_0)^{-1}$. Moreover, it follows from this condition and the continuity of $f_t(\theta_t)$ and $\Lambda(\theta_t)$ in t that $f_\tau(\theta_\tau) < \Lambda(\theta_\tau)^{-1}$. This inequality and the equilibrium condition $F_\tau(\theta_\tau) = I^{-1}(\theta_\tau)$ imply that F_τ crosses I^{-1} from below at θ_τ: that is,

$$\operatorname{sgn}(F_\tau(\theta) - I^{-1}(\theta)) = \operatorname{sgn}(\theta - \theta_\tau) \tag{7}$$

whenever $|\theta - \theta_\tau| \leq \varepsilon$. Additionally, since θ_t is continuous in t, there is an $\eta > 0$ such that $|\theta_t - \theta_\tau| \leq \varepsilon$ whenever $|t - \tau| \leq \eta$.

Let $v = \min(\operatorname*{arg\,min}_{s \in [\tau, \tau+\eta]} \theta_s)$. By the definition of τ, $v > \tau$, and by the definition of v, $\sigma(\alpha, \theta_t) = B$ for all $\alpha < \theta_v$ and $t \in [\tau, \tau + \eta]$. Thus, by Lemma A7,

$$f_v(\alpha) = f_\tau(\alpha)\exp\left(\int_\tau^v g_B(x_s)ds\right) \geq f_\tau(\alpha). \tag{8}$$

for all $\alpha < \theta_v$. It follows that

$$\begin{aligned} F_v(\theta_v) &= 1 - \int_{-\infty}^{\theta_v} f_v(\alpha)d\alpha && \text{by definition} \\ &\leq 1 - \int_{-\infty}^{\theta_v} f_\tau(\alpha)d\alpha && \text{by inequality (8)} \\ &= F_\tau(\theta_v) && \text{by definition} \\ &< I^{-1}(\theta_v) && \text{by equality (7).} \end{aligned}$$

Hence, $F_v(\theta_v) < I^{-1}(\theta_v)$, contradicting the equilibrium condition (E). ❑

The proof of Case 2 is completed in a fashion analogous to that of Case 1. ■

The following proposition establishes the laws of motion of θ_t and x_t.

Proposition A9: *Let* $\{f_t, \theta_t\}_{t\in[0,T)}$ *be a continuous solution to* (D) *and* (E). *Then*:

$$\dot{\theta}_t = \frac{x_t\, g_A(x_t)}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}} \quad and \quad \dot{x}_t = \frac{x_t\, g_A(x_t)}{1 - \Lambda(\theta_t) f_t(\theta_t)}. \qquad \text{(9) and (10)}$$

Proof: First, suppose that $\Lambda(\theta_0)\theta_0 \geq 0$. Lemmas A5 and A8 then imply that θ_t is increasing in *t*. Therefore, by Lemma A6, $f_t(\theta_t) = f_0(\theta_t)\gamma_t$ and $F_t(\theta_t) = F_0(\theta_t)\gamma_t$. Substituting these equations and condition (E) into equation (C1), we see that

$$\begin{aligned}\dot{\theta}_t &= \frac{F_0(\theta_t)\,\gamma_t\, g_A\big(I^{-1}(\theta_t)\big)}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}} \\ &= \frac{F_t(\theta_t)\, g_A\big(I^{-1}(\theta_t)\big)}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}} \\ &= \frac{x_t\, g_A(x_t)}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}}\end{aligned}$$

Since condition (E) implies that $x_t = I^{-1}(\theta_t)$ and since $\frac{d}{d\theta} I^{-1}(\theta_t) = -\Lambda(\theta_t)^{-1}$, we conclude that

$$\dot{x}_t = \tfrac{d}{d\theta} I^{-1}(\theta_t)\, \dot{\theta}_t = \frac{x_t\, g_A(x_t)}{1 - \Lambda(\theta_t) f_t(\theta_t)}.$$

Now suppose that $\Lambda(\theta_0)\theta_0 < 0$. In this case, an analysis similar to that used to prove Lemma A6 shows that $f_t(\theta_t) = f_0(\theta_t)\gamma_t$ and that $(1 - F_t(\theta_t)) = (1 - F_0(\theta_t))\gamma_t$. Therefore, the identity $x\, g_A(x) + (1 - x)\, g_B(x) \equiv 0$ implies that

$$\begin{aligned}\dot{\theta}_t &= \frac{\big(1 - F_0(\theta_t)\big)\,\gamma_t\, g_B\big(I^{-1}(\theta_t)\big)}{f_0(\theta_t)\,\gamma_t - \Lambda(\theta_t)^{-1}} \\ &= \frac{\big(1 - F_t(\theta_t)\big)\, g_B\big(I^{-1}(\theta_t)\big)}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}} \\ &= \frac{(1 - x_t)\, g_B(x_t)}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}} \\ &= \frac{x_t\, g_A(x_t)}{f_t(\theta_t) - \Lambda(\theta_t)^{-1}}.\end{aligned}$$

The proof is completed as before. ■

Proof of Theorem 1:

Existence, uniqueness, and continuity of solutions are proved using the ODE Reduction (Theorem A3). To prove existence, observe that since $\Lambda(\theta) < 0$ for all θ, equation (C1) is always well-defined; by assumption (A3) it is differentiable in θ_t and γ_t. Therefore, the existence of solutions to (C1) and (C2) on $[0, \infty)$ follows from standard results. Existence of solutions to (D) and (E) then follows from the implication (*ii*) ⇒ (*i*) of Theorem A3.

We now prove uniqueness and continuity. Proposition A1 (*ii*) and condition (D) guarantee that the component $\{\theta_t\}_{t\geq 0}$ of any solution to (D) and (E) is continuous in t. Hence, the implication (*i*) ⇒ (*ii*) of Theorem A3 tells us that any solution to (D) and (E), coupled with definition (G), must also solve (C1), (C2), and (C3). Since the solution to the latter set of equations is unique and continuous on $[0, \infty)$, so is the solution to the former set.

To characterize limit behavior, let $\{f_t, \theta_t\}_{t\geq 0}$ be the unique solution trajectory specified in Theorem A3, and let $x_t \equiv F_t(\theta_t)$. We consider the case in which $\theta_0 \leq 0$; the proof of the other case is similar. Lemma A5 tells us that θ_t is increasing over time. Therefore, since $x_t = I^{-1}(\theta_t)$ and $I^{-1}(\cdot)$ is increasing, $x_0 < x^*$ and x_t is increasing over time. Since g_A is a continuous function satisfying $\text{sgn}(g_A(x_t)) = \text{sgn}(x^* - x)$, x^* is a rest point of equation (10), and the right hand side of equation (10) is bounded below away from zero on any closed interval $[x_1, x_2] \subset (0, x^*)$. Therefore, by Proposition A9, $\lim_{t\to\infty} x_t = x^*$. Since $\theta_t = I(x_t)$ and since $I(x^*) = 0$ by definition, $\lim_{t\to\infty} \theta_t = 0$. ■

Proof of Theorem 2:

We present the proof for the case in which $\theta_0 > 0$; the proof when $\theta_0 < 0$ is similar, and the proof when $\theta_0 = 0$ is trivial.

We begin with a C^1 preference distribution f_0 and a bias threshold $\theta_0 > 0$ satisfying the equilibrium condition $\theta_0 = I(F_0(\theta_0))$ and the stability condition $f_0(\theta_0) < \Lambda(\theta_0)^{-1}$, which is implied by assumptions (A3), (A4), and (A5). The existence and uniqueness of an initial continuous solution trajectory from this initial condition is established using Theorem A3 just as in the proof of Theorem 1. Let $[0, T)$ be the maximal domain of the continuous solution trajectory: T is the latest time such that $\dot{\theta}_t < \infty$ for all $t < T$.

We proceed with five lemmas which characterize the initial conditions from which jumps occur. Together, these lemmas prove part (*ii*) of the theorem.

Lemma A10: *Suppose that θ_t is increasing through time T. Then if $\theta \in E_G(f_t) \cap [\theta_{t^-}, I(0)]$ and $t \in [0, T]$, then $f_t(\theta) = f^*(\theta)$. In particular, $f_t(\theta_t) = f^*(\theta_t)$ for all $t \in [0, T)$.*

Proof: Lemma A6 and the continuity of $f_{(\cdot)}(\alpha)$ and $\gamma_{(\cdot)}$ at T imply that for all $t \in [0, T]$ and $\theta \geq \theta_{t^-}$, $f_t(\theta) = f_0(\theta)\gamma_t$ and $F_t(\theta) = F_0(\theta)\gamma_t$. Moreover, since $\theta \in E_G(f_t)$, $F_t(\theta_t) = I^{-1}(\theta_t)$. Combining these three equations and the definition of f^* yields the result. ❑

Lemma A11: *If $T = \infty$, then $f^*(\theta) < \Lambda(\theta)^{-1}$ for all $\theta \geq \theta_0$, $\lim\limits_{t\to\infty} \theta_t = I(0)$, and $\lim\limits_{t\to\infty} x_t = 0$.*

Proof: If $T = \infty$, then $\dot{\theta}_t < \infty$ for all finite t. Thus, equation (9) implies that $f_t(\theta_t) < \Lambda(\theta_t)^{-1}$ for all $t \geq 0$. Therefore, since g_A is a continuous function satisfying $\mathrm{sgn}(g_A(x_t)) = \mathrm{sgn}(x - x^*)$, the right hand side of equation (10) is bounded below away from zero on any closed interval $[x_1, x_2] \subset (0, x^*)$. Thus, the limit of x_t cannot exceed 0. It follows that $\lim\limits_{t\to\infty} x_t = 0$ and that $\lim\limits_{t\to\infty} \theta_t = \lim\limits_{t\to\infty} I(x_t) = 0$. Moreover, since θ_t is increasing through time t and takes every value in $[\theta_0, I(0))$, Lemma A10 implies that $f^*(\theta)$ is less than $\Lambda(\theta)^{-1}$ on this interval, and therefore on $[\theta_0, \infty)$. ❑

If $T = \infty$, then the existence, uniqueness, continuity, and limit behavior of the solution trajectories follow from Theorem A3, the existence and uniqueness of solutions to ODEs, and Lemma A11. However, if $T < \infty$, θ_{T^-} is an unstable equilibrium under f_T, so a jump must occur at time T before continuous evolution can proceed. To establish that θ_{T^-} is unstable, we first prove:

Lemma A12: *If $T < \infty$, then $f_T(\theta_{T^-}) = \Lambda(\theta_{T^-})^{-1}$.*

Proof: If $T < \infty$, then by definition, $\lim\limits_{t\uparrow T} \dot{\theta}_t = \infty$. Therefore, equation (9) implies that $\lim\limits_{t\uparrow T} (f_t(\theta_t) - \Lambda(\theta_t)^{-1}) = 0$, from which the result follows. ❑

We now want to use Lemma A12 and Assumption (A5) to show that F_T and I^{-1} are strictly tangent at θ_{T^-}, and so that θ_{T^-} is an unstable equilibrium under f_T. To apply assumption (A5), we need to show that $f_T'(\theta_{T^-})$ exists.

Lemma A13: *$f_T(\cdot)$ is continuously differentiable on (θ_0, ∞).*

Proof: It is clear from equation (9) that θ_t is strictly increasing on [0, T). Therefore, the inverse of $\theta_{(\cdot)}$ exists. We call this inverse $\tau(\alpha)$ and define it as

$$\tau(\alpha) = \begin{cases} \theta^{-1}(\alpha) & \text{if } \alpha \in [\theta_0, \theta_{T^-}), \\ T & \text{if } \alpha = \theta_{T^-}. \end{cases}$$

Since $\theta(\cdot)$ is C^1 on (0, T), $\tau(\cdot)$ is C^1 on (θ_0, θ_{T^-}), and $\tau'(\alpha) = \left(\dot{\theta}_{\tau(\alpha)}\right)^{-1}$. Moreover, since $\lim\limits_{t \uparrow T} \dot{\theta}_t = \infty$,

$$\tfrac{d}{d\theta^-}\tau(\theta_{T^-}) = \lim_{\alpha \uparrow \theta_{T^-}} \tau'(\alpha) = \lim_{\alpha \uparrow \theta_{T^-}} \left(\dot{\theta}_{\tau(\alpha)}\right)^{-1} = \lim_{t \uparrow T}\left(\dot{\theta}_t\right)^{-1} = 0.$$

Because θ_t increases over time, and since each $f_{(\cdot)}(\alpha)$ is continuous by condition (D),

$$f_t(\theta) = \begin{cases} f_0(\theta)\exp\left(\int_0^{\tau(\theta)} g_A(x_s)ds + \int_{\tau(\theta)}^t g_B(x_s)ds\right) & \text{if } \theta_0 \le \theta < \theta_{t^-}, \\ f_0(\theta)\exp\left(\int_0^t g_A(x_s)ds\right) & \text{if } \theta \ge \theta_{t^-}. \end{cases}$$

for all $t \in$ [0, T]. Therefore, for all t in this range,

$$\tfrac{d}{d\theta} f_t(\theta) = \begin{cases} \left(\tfrac{d}{d\theta} f_0(\theta)\right)\exp\left(\int_0^{\tau(\theta)} g_A(x_s)ds + \int_{\tau(\theta)}^t g_B(x_s)ds\right) + \left(\tfrac{d}{d\theta}\tau(\theta)\right)f_0(\theta) & \\ \quad \times (g_A(x_{\tau(\theta)}) - g_B(x_{\tau(\theta)}))\exp\left(\int_0^{\tau(\theta)} g_A(x_s)ds + \int_{\tau(\theta)}^t g_B(x_s)ds\right) & \text{if } \theta_0 < \theta < \theta_{t^-}, \\ \left(\tfrac{d}{d\theta} f_0(\theta)\right)\exp\left(\int_0^t g_A(x_s)ds\right) & \text{if } \theta > \theta_{t^-}. \end{cases}$$

Since f_0 is C^1 by assumption (A3), this proves the result for $\theta \neq \theta_{T^-}$. To complete the proof, observe that $\tfrac{d}{d\theta^-} f_t(\theta_{t^-})$ is defined analogously to the first case above and $\tfrac{d}{d\theta^+} f_t(\theta_{t^-})$ to the second, substituting one-sided derivatives wherever appropriate. Thus, since $\tfrac{d}{d\theta^-}\tau(\theta_{T^-}) = 0$,

$$\begin{aligned} \tfrac{d}{d\theta^-} f_T(\theta_{T^-}) &= \tfrac{d}{d\theta^-} f_0(\theta_{T^-}) \exp\left(\int_0^T g_A(x_s)ds\right) \\ &\quad + \left(\tfrac{d}{d\theta^-}\tau(\theta_{T^-})\right)f_0(\theta_{T^-})(g_A(x_{T^-}) - g_B(x_{T^-})) \exp\left(\int_0^T g_A(x_s)ds\right) \\ &= \tfrac{d}{d\theta} f_0(\theta_{T^-}) \exp\left(\int_0^T g_A(x_s)ds\right) \\ &= \tfrac{d}{d\theta^+} f_T(\theta_{T^-}). \quad ❑ \end{aligned}$$

Lemma A13 and assumption (A5) guarantee that F_T and I^{-1} are tangent at θ_{T^-}, and hence that θ_{T^-} is unstable. The following lemma shows that the former curve lies below the latter near θ_{T^-}, as pictured in Figure 8.

Lemma A14: *$F_T(\theta) < I^{-1}(\theta)$ for all $\theta \in (\theta_0, \theta_{T^-})$.*

Proof: Fix $\theta \in (\theta_0, \theta_{T^-})$. Since θ_t is continuous and strictly increasing on $[0, T)$ with range $[\theta_0, \theta_{T^-})$, there is a unique time $t \in (0, T)$ such that $\theta_t = \theta$; at all times $s \in (t, T)$, players whose bias is less than θ use strategy B. Lemma A7 implies that $f_T(\alpha) > f_t(\alpha)$ for all $\alpha < \theta$, so

$$F_T(\theta) = 1 - \int_\theta^\infty f_T(\alpha)\, d\alpha < 1 - \int_\theta^\infty f_t(\alpha)\, d\alpha = F_t(\theta) = I^{-1}(\theta),$$

where the final equality follows from condition (E). ❑

Proposition A15 establishes that the jump from θ_{T^-} is well defined. Its proof uses three new definitions. Let $N_\varepsilon(\theta) = (\theta - \varepsilon, \theta + \varepsilon) - \{\theta\}$ be the ε-neighborhood of θ, and let $N_\varepsilon^-(\theta) = (\theta - \varepsilon, \theta)$ and $N_\varepsilon^+(\theta) = (\theta, \theta + \varepsilon)$ be the parts of the ε-neighborhood which lie to the left and right of θ, respectively.

Recall that when $\theta_{T^-} > 0$, the jump function is given by

$$\theta_T = J(f_T, \theta_{T^-}) = \min \{\theta > \theta_{T^-} : \theta \in SE_G(f_T)\}.$$

Proposition A15: *$J(f_T, \theta_{T^-})$ is well defined.*

Proof: The first step in the proof is an inductive argument which considers the elements of $E_G(f_T) \cap [\theta_{T^-}, I(0)]$ in increasing order. The inductive hypothesis is as follows: Suppose that $\theta^k \in E_G(f_T) \cap [\theta_{T^-}, I(0)]$, that $f_T(\theta^k) = \Lambda(\theta^k)^{-1}$, and that $F_T(\theta) < I^{-1}(\theta)$ on $N_\varepsilon^-(\theta^k)$ for some $\varepsilon > 0$. Then $\theta^{k+1} = \min E_G(f_T) \cap (\theta^k, I(0)]$ is well-defined, and $F_T(\theta) < I^{-1}(\theta)$ on $N_\eta^-(\theta^{k+1})$ for some $\eta > 0$. We begin the induction at $\theta^0 = \theta_{T^-}$; the induction ends once $f_T(\theta^k) \neq \Lambda(\theta^k)^{-1}$. It is clear that this process moves through the elements of $E_G(f_T) \cap [\theta_{T^-}, I(0)]$ in ascending order. We show below that the process terminates in a finite number of steps, and that this implies that $J(f_T, \theta_{T^-})$ is well defined and equal to the final value of θ^k.

Lemmas A12 and A14 provide the basis for the induction. We now show that the inductive hypothesis is true. To begin, we show that under the conditions of the inductive hypothesis, F_T and I^{-1} are strictly tangent at the point θ^k. Since $\theta^k \in E_G(f_T)$, $F_T(\theta^k) = I^{-1}(\theta^k)$; that $f_T(\theta_{T^-}) = \Lambda(\theta_{T^-})^{-1}$ and the terminal condition of the induction imply that $F_T'(\theta^k) = -f_T(\theta^k) = -\Lambda(\theta^k)^{-1} = (I^{-1})'(\theta^k)$. Thus, to establish the strict tangency it is sufficient to show that $F_T''(\theta^k) < (I^{-1})''(\theta^k)$.

Choose $\theta \in N_\varepsilon^-(\theta^k)$. Taking the Taylor expansion of $I^{-1}(\theta) - F_T(\theta)$ about θ^k reveals that

$$\begin{aligned} I^{-1}(\theta) - F_T(\theta) &= (I^{-1}(\theta^k) - F_T(\theta^k)) + ((I^{-1})'(\theta^k) - F_T'(\theta^k))(\theta - \theta^k) \\ &\quad + \tfrac{1}{2}((I^{-1})''(\tilde{\theta}) - F_T''(\tilde{\theta}))(\theta - \theta^k)^2 \\ &= \tfrac{1}{2}((I^{-1})''(\tilde{\theta}) - F_T''(\tilde{\theta}))(\theta - \theta^k)^2, \end{aligned}$$

where $\tilde{\theta} \in (\theta, \theta^k)$. Since $F_T(\theta) < I^{-1}(\theta)$ by assumption, $((I^{-1})''(\tilde{\theta}) - F_T''(\tilde{\theta}))(\theta - \theta^k)^2 > 0$. Letting θ increase to θ^k, we find that $(I^{-1})''(\theta^k) - F_T''(\theta^k) \geq 0$. Therefore, the conditions of the inductive hypothesis, Lemma A13, and assumption (A5) imply that $(I^{-1})''(\theta^k) - F_T''(\theta^k) > 0$. Hence, F_T and I^{-1} are strictly tangent at θ^k.

Because of this strict tangency, $I^{-1}(\theta) > F_T(\theta)$ on $N_\delta(\theta^k)$ for some $\delta > 0$. Thus, $E_G(f_T) \cap (\theta^k, I(0)] = E_G(f_T) \cap [\theta^k + \delta, I(0)]$, the latter of which is the intersection of two compact sets. Therefore, θ^{k+1}, which is the minimum of this set, is well-defined: it is the first equilibrium to the right of θ^k. Finally, since F_T and I^{-1} are continuous functions, $F_T(\theta) < I^{-1}(\theta)$ on $N_\eta^-(\theta^{k+1})$ for some $\eta > 0$. This completes the proof of the inductive step.

We now establish that the inductive process terminates in a finite number of steps. Suppose to the contrary that it does not. Then since the θ^k form an increasing sequence which is bounded above by $I(0)$, their limit $\bar{\theta}$ exists. Since F_T and I are continuous, $F_T(\bar{\theta}) = \lim_{k\to\infty} F_T(\theta^k) = \lim_{k\to\infty} I^{-1}(\theta^k) = I^{-1}(\bar{\theta})$, and so $\bar{\theta} \in E_G(f_T)$; since f_T and Λ are continuous, $f_T(\bar{\theta}) = \lim_{k\to\infty} f_T(\theta^k) = \lim_{k\to\infty} \Lambda(\theta^k)^{-1} = \Lambda(\bar{\theta})^{-1}$. Furthermore, since $(I^{-1})''(\theta) - F_T''(\theta)$ is continuous on $[\theta_{T^-}, I(0)]$ and strictly positive for each θ^k, $(I^{-1})''(\bar{\theta}) + F_T''(\bar{\theta}) \geq 0$. If we substitute $\bar{\theta}$ for θ^k, the argument two paragraphs above shows that $(I^{-1})''(\bar{\theta}) + F_T''(\bar{\theta}) > 0$. Thus, F_T and I^{-1} are strictly tangent at $\bar{\theta}$: $F_T(\theta) < I^{-1}(\theta)$ on $N_\kappa(\bar{\theta})$ for some $\kappa > 0$. However, this last statement contradicts that $\bar{\theta}$ is an accumulation point of the sequence of equilibria $\left\{\theta^k\right\}_{k=0}^{\infty}$. Therefore, the inductive process must terminate in a finite number of steps, arriving at θ_T. We have thus established that θ_T is well-defined and that $F_T(\theta) \leq I^{-1}(\theta)$ on $[\theta_{T^-}, \theta_T]$.

To complete the proof, we must show that $f_T(\theta_T) < \Lambda(\theta_T)^{-1}$. Since θ_T is the final point reached by the inductive process, we know that and $F_T(\theta) < I^{-1}(\theta)$ on $N_\eta^-(\theta_T)$ for some $\eta > 0$. Choosing $\theta \in N_\eta^-(\theta_T)$ and taking the Taylor expansion of $I^{-1}(\theta) - F_T(\theta)$ about θ_T, we find that

$$\begin{aligned} I^{-1}(\theta) - F_T(\theta) &= (I^{-1}(\theta_T) - F_T(\theta_T)) + ((I^{-1})'(\tilde{\theta}) - F_T'(\tilde{\theta}))(\theta - \theta_T) \\ &= (-\Lambda(\tilde{\theta})^{-1} + f_T(\tilde{\theta}))(\theta - \theta_T) \end{aligned}$$

for some $\tilde{\theta} \in (\theta, \theta_T)$. Therefore, $f_T(\tilde{\theta}) < \Lambda(\tilde{\theta})^{-1}$. Letting θ approach θ_T from below, we see that $f_T(\theta_T) \leq \Lambda(\theta_T)^{-1}$. But $f_T(\theta_T) \neq \Lambda(\theta_T)^{-1}$ by the terminal condition of the induction; hence, $f_T(\theta_T) < \Lambda(\theta_T)^{-1}$. ❑

The complete solution trajectory $\{f_t, \theta_t\}_{t \geq 0}$ to (D), (E), and (C) begins with an initial round of continuous evolution. Alternating rounds of jumps and continuous evolution follow through the infinite time horizon; the number of distinct rounds may or may not be finite. Theorem A3 guarantees that each round of continuous evolution is uniquely defined given its initial conditions; Proposition A15 does the same for jumps. Together, these results establish the existence and uniqueness of the entire solution trajectory. This establishes part (*i*) of the theorem.

All that remains is to characterize limit behavior. To do so, we consider the trajectory of the strategy distribution x_t. At each instant t, the strategy distribution is either adjusting downward at a rate given by equation (10) or is making a discrete jump downward as dictated by the jump function J. Moreover, since g_A is a continuous function satisfying $\operatorname{sgn}(g_A(x_t)) = \operatorname{sgn}(x - x^*)$, the right hand side of equation (10) is bounded below on any closed interval $[x_1, x_2] \subset (0, x^*)$. Consequently, the limit of x_t cannot exceed 0. We therefore conclude that $\lim_{t\to\infty} x_t = 0$ and that $\lim_{t\to\infty} \theta_t = \lim_{t\to\infty} I(x_t) = I(0)$. This completes the proof of Theorem 2. ■

Proof of Corollary 3:

To prove the first claim, we observe that if $f_0(\theta) < \min_{\theta \in [I(1), I(0)]} \Lambda(\theta)^{-1}$, then F_0 and I^{-1} intersect exactly once, and their intersection is the initial equilibrium θ_0. Furthermore, $\operatorname{sgn}(F_0(\theta) - I^{-1}(\theta)) = \operatorname{sgn}(\theta - \theta_0)$ for all θ in the range of I. The claim then follows immediately from the definition of f^* and Theorem 2 (*ii*).

To prove the second claim, consider the case in which $\theta_0 > 0$; the proof when $\theta_0 < 0$ is similar. Since $\theta_0 \in E_G(f_0)$, $F_0(\theta_0) = I^{-1}(\theta_0)$; assumption (A3) requires that $f_0(\theta_0) <$

$\Lambda(\theta_0)^{-1}$. Therefore, taking a Taylor expansion of $F_0(\theta) - I^{-1}(\theta)$ about θ_0 reveals that $F_0(\theta) > I^{-1}(\theta)$ for all $\theta \in N_{\varepsilon}^{+}(\theta_0)$ for some $\varepsilon > 0$: at $t = 0$ there are no equilibria in a neighborhood to the right of θ_0. Therefore, since $E_G(f_0)$ is closed and $E_G(f_0) \cap (\theta_0, I(0))$ is non-empty, there is a first equilibrium to the right of θ_0, which we denote $\theta^* = \min E_G(f_0) \cap (\theta_0, I(0))$.

We now show that $f_0(\theta^*) \geq \Lambda(\theta^*)^{-1}$. Suppose to the contrary that $f_0(\theta^*) < \Lambda(\theta^*)^{-1}$. Then, since $\theta^* \in E_G(f_0)$, $F_0(\theta^*) = I^{-1}(\theta^*)$, and so $F_0(\theta) < I^{-1}(\theta)$ for all $\theta \in N_{\delta}^{-}(\theta^*)$ for some $\delta > 0$. Therefore, the Intermediate Value Theorem guarantees that $F_0(\theta) = I^{-1}(\theta)$ for some $\theta \in (\theta_0, \theta^*)$, contradicting that θ^* is the first equilibrium to the right of θ_0.

Since θ_0 and θ^* are equilibria under f_0, $f^*(\theta^*) = f_0(\theta^*) \geq \Lambda(\theta^*)^{-1}$ and $f^*(\theta_0) = f_0(\theta_0) < \Lambda(\theta_0)^{-1}$ by Lemma A10. Therefore, applying the Intermediate Value Theorem to $f_0(\cdot) - \Lambda(\cdot)^{-1}$ establishes that $f^*(\theta) = \Lambda(\theta)^{-1}$ for some $\theta \in (\theta_0, \theta^*]$. Theorem 2 (*ii*) then guarantees that a jump must occur. ■

References

Bowles, S. (1998). “Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions.” *Journal of Economic Literature* **36**, 75-111.

Cooper, W. S. (1987). “Decision Theory as a Branch of Evolutionary Theory: A Biological Derivation of the Savage Axioms.” *Psychological Review* **4**, 395-411.

Dekel, E., J. C. Ely, and O. Yilankaya (1998). "Evolution of Preferences." mimeo, Northwestern University.

Dekel, E., and S. Scotchmer (1999). “On the Evolution of Attitudes Towards Risk in Winner-Take-All Games.” *Journal of Economic Theory* **87**, 125-143.

Durlauf, S. N. (1997). "Statistical Mechanics Approaches to Socioeconomic Behavior." In *The Economy as an Evolving Complex System II*, W. B. Arthur, S. N. Durlauf, and D. A. Lane, Eds. Reading, MA: Addison-Wesley.

Ely, J. C., and O. Yilankaya (1997). “Nash Equilibrium and the Evolution of Preferences.” CMS-EMS Discussion Paper #1191, Northwestern University.

Güth, W. (1995). “An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives.” *International Journal of Game Theory* **24**: 323-344.

Güth, W., and M. Yaari (1992). “Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach.” In U. Witt, ed., *Explaining Process and*

Change: *Approaches to Evolutionary Economics*, 23-34. Ann Arbor: University of Michigan Press.

Harsanyi, J. C. (1973). “Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points.” *International Journal of Game Theory* **2**, 1-23.

Huck, S., G. Kirchsteiger, and J. Oechssler (1999). "Learning to Like What You Have: Explaining the Endowment Effect." mimeo, Humboldt University Berlin and University of Vienna.

Huck, S., and J. Oechssler (1999). "The Indirect Evolutionary Approach to Explaining Fair Allocations." *Games and Economic Behavior* **28**, 13-24.

Kandori, M., G. J. Mailath, and R. Rob (1993). "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica* **61**, 29-56.

Karni, E., and D. Schmeidler (1986). “Self-preservation as a Foundation of Rational Behavior under Risk.” *Journal of Economic Behavior and Organization* **7**, 71-82.

Kuran, T. (1989). "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution," *Public Choice* **61**, 41-74.

Kuran, T. (1991). "The East European Revolution of 1989: Is It Surprising that We Were Surprised?" *American Economic Review Papers and Procedings* **81**, 121-125.

Kuran, T. (1995). *Private Truths, Public Lies*. Cambridge: Harvard University Press.

Poston, T., and I. Stewart (1978). *Catastrophe Theory and Its Applications*. London: Pitman.

Robson, A. J. (1996a). “A Biological Basis for Expected and Non-expected Utility.” *Journal of Economic Theory* **68**, 397-424.

Robson, A. J. (1996b). “The Evolution of Attitudes Towards Risk: Lottery Tickets and Relative Wealth.” *Games and Economic Behavior* **14**, 190-207.

Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge: MIT Press.