Is Inequality an Evolutionary Universal?

Samuel Bowles

1. Introduction[1]

Hernán Cortés' long letters to King Charles of Castile describe the exotic and unusual customs he and his armed band encountered as they advanced toward Temixtitan in 1519. But in light of the thirteen millennia or more that had passed since there had been any sustained contact between people of the Old World and the New, what is striking about his account of Mexico is how familiar it all was. Upon reaching Temixtitan (modern day Mexico City), he wrote:

> *There are many chiefs, all of whom reside in this city, and the country towns contain peasants who are vassals of these lords and each of whom holds his land independently; some have more than others.. And there are many poor people who beg from the rich in the streets as the poor do in Spain and in other civilized places.* (Cortes (1986):68, 75)

He remarks also that "the orderly manner which, until now, these people have been governed is almost like that of the states of Venice or Genoa or Pisa."

Some scholars, like Cortés, are impressed by the similarity of institutions in quite differing

[1] Thanks to Ugo Pagano, Yong-jin Park, Bae Smith, Elisabeth Wood, Jorgen Weibull and Peyton Young for valuable contributions to this paper, and to the PEW Charitable Trust and the Santa Fe Institute for financial support. The formal mathematical analysis of the processes described here can be found in Bowles (2004). An earlier version appears in Bowles, Durlauf, and Hoff (2004)

environments and have postulated a set of social arrangements that are favored by historical evolutionary processes. Talcott Parsons (1964) termed these *evolutionary universals*, namely, those ways of ordering society which crop up with sufficient frequency in a variety of circumstances to suggest their general evolutionary viability. Parsons offered vision as a biological analogy to these evolutionary universals, another biological example would be sexual reproduction; both have emerged under a wide variety of circumstances and in a great many species. Among the human examples that Parsons identifies as evolutionary universals are money, bureaucracy, and "social stratification."

Are inegalitarian institutions evolutionary universals? Some types of inegalitarian social arrangements seem to be highly persistent and yet to offer no advantages in productive efficiency. Leading examples are the monopolization of political rights by elites, the caste system, the exclusion of all but the well to do from entrepreneurial pursuits, and blocked access to basic education and medical care. The purpose of this paper is to see what light evolutionary theory coupled with the theory of collective action can shed on such questions as: *Why have institutions that implement highly unequal divisions of the social product been so ubiquitous since the domestication of animals and plants 11 millennia ago? Why do they persist when they convey no clear efficiency advantages over other feasible social arrangements*?

Institutions that implement widespread poverty and that persist over long periods despite their lack of productive superiority over alternative more egalitarian institutions are what I term *institutional poverty traps*. Some institutional poverty traps implement low average incomes for the populations which they affect.

An example, due to Banerjee and Iyer (2002), also illustrates the way that institutions can

induce poor aggregate economic performance with poverty as its correlate. One effect of British rule in India was a growing entrenchment of the power and property rights of powerful landlords. Their influence was already substantial under the Mughal rulers before the British, but during the Bengal Presidency it was greatly strengthened by the Permanent Settlement of 1793. This act of the colonial rulers conferred *de facto* governmental powers to the landlords (the *zamindars*) by giving them the right to collect taxes (keeping a substantial fraction for themselves). The fact that British taxation and land tenure policy was not uniform throughout the Raj provides a natural experiment to test the importance of institutions. Banerjee and Iyer compared the post-Independence economic performance and social indicators of districts of modern-day India in which landlords had been empowered by the colonial land tenure and taxation systems with other districts where the landlords had been bypassed in favor of the village community or direct taxation of the individual cultivator. They found that the landlord-controlled districts had significantly *lower* rates of agricultural productivity growth stemming from lower rates of investment and lesser use of modern inputs. The landlord-controlled districts also lagged significantly in educational and health improvements.[2]. These findings suggest that enduring poverty can result from the persistence of an institutional innovation occurring a century or more earlier.

The institutional poverty traps studied by Banerjee and Iyer and Engerman and Sokolof

[2] The causal connection between landlord control and these subsequent results remains to be explored. Because colonial practices changed over time in response to events such as the revolt by Indian soldiers in 1857 and over space in response to the idiosyncracies of local administrators, Banerjee and Iyer were able to identify independent sources of variation in the land tenure and taxation policies not due to pre-existing conditions.

concern the ways that institutions result in low levels of *aggregate* economic performance: even the relatively well to do in the *zamindari* regions of India are not well off. Here I will investigate another type of institutional poverty trap namely those under which poverty results due to the unequal division of output between classes -- that is, individuals with different structural positions in the economy, such as landlords and share croppers or employers and workers. Examples of such divisions include social structures in which a single landlord claims half or more of the crop of ten or more tenants, thus receiving or order of magnitude more income than the average tenant while putting in substantially less effort. Another example is the class structure of the classical capitalist firm, in which an owner pays wages equivalent to two thirds of the value added produced by each of a hundred or more workers he employs, resulting in an income difference of 50:1. An important question to be explored is: does the persistence of these highly unequal divisions require that the associated institutions be productively superior in some sense to alternative social arrangements? Or may their persistence be explained by other evolutionary advantages they possess?

In these two examples, an institution may be represented as one of a number of possible conventions, that is, equilibria in which members of a population typically act in ways that maximize payoffs given the actions taken by others and in which individuals' beliefs about what other will do support continued adherence to these conventional actions. A convention is thus an outcome in which it is in the interest of people to adhere to the convention as long as they believe that most others will do the same. This property is what justifies the term "mutual best response" to describe a convention. A common non economic example of a convention is driving on the right or on the left. Examples of distributional conventions include simple principles of division such as "finders keepers" or "first come first served," as well as more complicated principles of allocation such as the

variety of rules which have governed the exchange of goods or the division of the products of one's labor over the course of human evolution.

Because a convention is one of many possible mutual best responses, institutions are not environmentally determined, but rather are of human construction (but not necessarily of deliberate design). Conventions are self-enforcing (that is they are evolutionarily stable) as a result of the positive feedbacks associated with the members' conforming to a common strategy. As a result of this, institutions that take the form of conventions will display inertia, and transitions among them will occur rapidly but infrequently, displaying a pattern that biologists call *punctuated equilibria*. (Eldredge and Gould (1972)).

To explore why institutional poverty traps are seemingly common, we need to understand the birth, diffusion, and eclipse of institutions and the process by which one institution supplants another. This will require an account of how characteristics of institutions contribute to their evolutionary success. Institutions may proliferate as the result of the ability of the groups governed by them to be victorious in wars, or to survive ecological crises when other perish. But to some extent the evolutionary success of institutions depends on how likely a group is to "hit upon" differing social arrangements, and, once hit upon, how resilient an institution is to pressures for change from the members of the group. Here I will set aside the effects of between group competition and differential group survival to focus on the within group processes that make some institutions easy to "hit upon" and resilient once in place. Two quite distinct approaches to the within-group processes bringing about institutional innovation may be identified.

The first, similar to the biologist Sewall Wright's use of drift to explain a movement from one fitness peak across a fitness valley to another peak (Wright (1931)), is that proposed by

stochastic evolutionary game theory pioneered by Foster and Young (1990). In this Darwin-inspired approach, change occurs through the chance bunching of individuals' mutation-like idiosyncratic behavior. If in a given period sufficiently many people do not adhere to the convention, then adhering to the convention will no longer be a best response, and so the convention will unravel. If this happens, a new convention may be established, much as a neighborhood may "tip" from being mostly of one race to mostly of another. Changes in language use, the racial composition of neighborhoods, contractual shares, market days, and etiquette have been modeled in this manner

The second approach, initiated by Marx, stresses asymmetries among the players and explains institutional innovation by the changing power balance between those who benefit from differing conventions. In this framework, revolutionary change in institutions is likely when existing institutions facilitate the collective action of those who would benefit from a change in institutions, and when, because existing institutions are inefficient by comparison to an alternative, there are substantial potential gains to making a switch. This collective-action-based approach has been used to model conflicts among classes resulting in a basic transformation of social organization, such as the French, Russian, and Cuban revolutions as well as more gradual changes in institutional arrangements such as the centuries-long erosion of European feudalism.

Do these approaches allow us to say anything about the characteristics of evolutionarily successful institutions? Though the underlying causal mechanisms are different, the Marx-inspired approach shares with Darwin-inspired stochastic evolutionary game theory the prediction that institutional arrangements which are both inefficient and highly unequal will bear an evolutionary disability and will tend to be displaced in the long run by more efficient and more egalitarian

institutions.[3] This is quite an arresting claim in light of the long-term historical persistence of social arrangements which would appear to be neither efficient nor egalitarian. But the only formal model supporting this claim – the stochastic evolutionary game theory model due to Young (1998) – may be of limited relevance to real historical processes. The reason is that the "tipping" events that alter institutions are not generally induced by mutation-like accidents of behavior, but are rather the result of the intentional collective action of people who are not making mistakes, but rather are trying to better their condition.

Thus to address questions of institutional persistence, evolutionary theory must be coupled with the theory of collective action. A plausible account of collective action, combined with the fact that the classes are often of quite different sizes – many sharecropers, few landlords, for example – leads to a conclusion at odds with that of stochastic evolutionary game theory. This is that while more efficient and more equal institutions are indeed favored by plausible evolutionary processes under some conditions, it is also true that inefficient and unequal institutions can persist over very long periods of time.

2. Evolutionary Models of Institutional Transformation

Structural and social engineering approaches to the rise, persistence and fall of institutions, study constitution-making as a deliberate process whereby powerful groups or farsighted planners implement the rules that govern social interactions, seeking thereby to further their own or the public's interest. By contrast, evolutionary approaches represent institutions as an emergent property arising from the uncoordinated actions of members of a population, of which none are seeking to

[3] Efficient institutions yield a larger joint surplus, while in a more equal convention, the share of the least well-off is larger.

implement aggregate outcomes. The genes accounting for the design of the bird's wing were not trying to fly, but only to replicate, and in like manner, the actions accounting for many customs, norms, and other aspects of social interaction are represented in evolutionary models as resulting from the actions of individuals motivated by other goals, these structures being unintended byproducts. The approach developed below is evolutionary in spirit and formal methods, but it incorporates the fact that when people act in ways that change institutions they were often attempting to do just that, even if the resulting processes do not reflect their objectives.

Because nothing of importance concerning the main points below is lost in taking an especially simple case, I confine myself to the analysis of the evolutionary dynamics governing transitions between two conventions in a two-person two-strategy game in a large population of individuals subdivided into two groups, the members of which are randomly paired to interact in a non-cooperative game with members of the other group.

Individuals' best- response play is based on a single-period memory, and they maximize their expected payoffs based on the distribution of the population in the previous period. The two population subgroups, initially assumed to be of equal size, are termed A's and B's, and each when paired with a member of the other group may chose action 1 or 0, with the A's payoffs, a_{ij} representing the payoff to an A-person playing action *i* against a B-person playing action *j*, and analogously for the B's. If the members of the pair choose the same action they get positive benefits, while if they chose different

Figure 1
Payoffs in the contract game

	B offer contract 1	B offer contract 0
A offer contract 1	$a_{11}b_{11}$	0 0
A offer contract 0	0 0	a_{00} b_{00}

actions they get nothing. For concreteness, suppose the sub-groups are economic classes selecting a contract to regulate their joint production, which will only take place if they agree on a contract. Payoffs are shares of the joint surplus of the project, with the no-production outcome normalized to zero for both. The payoffs, with the A's as the row player, and the B's as column player, are given in Figure 1.

To capture the conflict of interest between the two groups, let us assume that $b_{00} > b_{11} = a_{11} > a_{00} > 0$ so the B's strictly prefer the outcome in which both play 0 and the A's prefer the equal division outcome which results when both play 1.[4] Both of these outcomes are strict Nash equilibria, and thus both represent conventions, which I will denote E_0 and E_1 (or {0,0} and {1,1}). Both populations are normalized to unit size, so I refer equivalently to the numbers of players and the fraction of the population, abstracting from integer problems.

The state of this population in any time period t is $\{\alpha_t,\beta_t\}$, where α is the fraction of the A's who played 1 in the previous period and β is the fraction of the B's who played 1. For any state of the population, expected payoffs a_i and b_i for the A's and B's respectively playing strategy i, depend on the distribution of play among the opposing group in the previous period, or dropping the time subscript:

$a_1 = \beta a_{11}$; $a_0 = (1-\beta)a_{00}$; $b_1 = \alpha b_{11}$; and $b_0 = (1-\alpha)b_{00}$.

The relationship between the population state and the expected payoffs to each action is illustrated

[4] I refer to {1,1} as the "equal" convention as a shorthand. The levels of well-being attained by the A's and B's cannot be determined without additional information. (If the A's are share croppers who interact with only one B (a landlord), while B's interact with many A's, the "equal" convention would exhibit unequal incomes of the two groups, for example.)

in Figure 2.

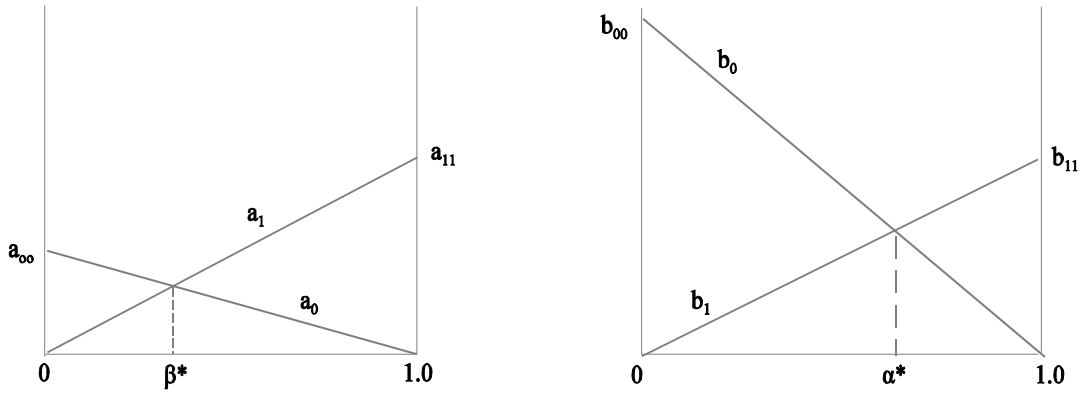

Figure 2 Expected payoffs depend on the distribution of play in the previous period.. Note: A's payoffs depend on β the fraction of B's offering contract 1, while the B's payoffs depend on α the fraction of A's offering contract 1. Because $b_{00} > b_{11} = a_{11} > a_{00}$, the convention E_1 (that is, α=1=β) is preferred by the A's while E_0 is preferred by the B's.

Individuals take a given action -- they are 1-players or 0-players -- and they continue doing so from period to period until they update their action, at which point they may switch. Suppose that at the beginning of every period some fraction ω of each sub-population may update their actions. (This might be due to the age structure of the population, with updating taking place only at a given period of life, in which case the "periods" in the model may be understood as "generations". Of course, updating could be much more frequent.) The updating is based on the expected payoffs to the two actions; these expectations are simply the payoffs which would obtain if the previous period's state remained unchanged (the population composition in the previous period being common knowledge in the current period.) While this updating process is not very sophisticated, it may realistically reflect individuals' cognitive capacities and it assures that in equilibrium -- when the

population state is stationary -- the beliefs of the actors formed in this naive process are confirmed in practice.

I will analyze the single-period change in the population state ($\Delta\alpha,\Delta\beta$) under the assumption that individual updating of strategies is monotonic in average payoffs so that $\Delta\alpha$ and $\Delta\beta$ have the signs respectively, of (a_1 - a_0) and (b_1 - b_0). The resulting population dynamics are illustrated in Figure 3, where the relevant regions are defined by:

$$\alpha^* = b_{00}/(b_{11}+b_{00})$$

$$\beta^* = a_{00}/(a_{11}+a_{00})$$

these two population distributions equating the expected payoffs to the two strategies for the two sub-populations, respectively. These values of α and β define best response functions: for $\alpha<\alpha^*$ B's best response is to play 0, and for $\alpha\geq\alpha^*$ B's best response is to play 1, with β^* interpreted analogously.

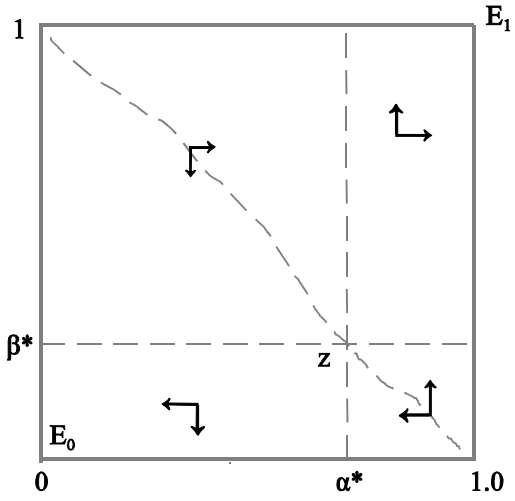

Figure 3 The state space. The arrows give the direction of change at every point. Note: E_1 and E_0 are stable equilibria; z is a saddle.

For states $\alpha<\alpha^*$ and $\beta<\beta^*$ (in the southwest region of Figure 3) it is obvious that $\Delta\alpha$ and $\Delta\beta$ are both negative and the state will move to {0,0}. Analogous reasoning holds for the northeast region. In the northwest and southeast regions of the state space we may define a locus of states from which the system will transit to the interior equilibrium α^*,β^*, with states below that locus transiting to {0,0}, and above the locus to {1,1}. The area below the dashed downward-sloping line in Figure 3 is the basin of attraction of {0,0} if a

population governed by the above dynamic finds itself in this region it will move to the {0,0} convention. The size of the basin of attraction of {0,0} varies with $\alpha^*\beta^*$. While the interior equilibrium $\{\alpha^*,\beta^*\}$ is an unstable Nash equilibrium (a saddle), the outcomes {0,0} and {1,1} are absorbing states of the dynamic process, meaning that if the population is ever at either of these states, it will never leave. There being more than one such absorbing state, the dynamic process is *non-ergodic*, that is, its long-run average behavior is dependent on initial conditions.

The critical values α^* and β^* will be important in what follows. Suppose the reigning convention is {0,0} and the A's would like to induce a shift to {1,1}, which is more favorable to them. If some fraction of the A's greater than α^* were to "strike" refusing to accept a_{00} and insisting on that a_{11} instead, then figure 2 confirms that in the next period the B's, having encountered these idiosyncratic A's would best respond by switching to the 1-contract, thereby tipping the population from {0,0} to {1,1}. Thus α^* is a "collective action threshold" indicating the minimum number of deviant A's required to displace the {0,0} equilibrium. It is easy to see (from figure 2) that were the {0,0} convention to become even more unequal (by increasing b_{00}) the effect would be to increase α^*, thereby raising the collective action threshold for the A's.

Similar reasoning holds for the B's of course. If the status quo convention were {1,1}, the B's might wish to induce a shift to {0,0}, which they could do if $(1-\beta^*)$ of them deviated from the best response "locking out" any A's unwilling to transact according the 0-contract. The effect on the B's collective action threshold of the 0-contract becoming even more unequal (lowering a_{00}) can also be seen from figure 2: it lowers β^* and hence makes it more difficult for the B's to induce a tipping

event to bring about a shift to the convention they would prefer.

Figure 4
Modified payoffs in the contract game

	B offer Contract 1	B offer Contract 0
A offer Contract 1	a_{11}=1 b_{11}=1	0 0
A offer Contract 0	0 0	a_{00}=σρ, b_{00}=(1-σ)ρ

The important point here is that were the 0-contract more unequal, it would make it more difficult for the A's to induce a shift away from this convention to the {1,1} convention they prefer; but it also makes it more difficult for the B's to induce a shift away from the 1-contract to the {0,0} convention that they prefer. The result is that increasing the inequality of the more unequal convention makes it more likely that the population will be 'stuck' in one convention or the other. But we cannot say whether greater inequality in of the {0,0} convention will result in the population being more likely in the long run to be there as opposed to the {1,1} convention. To do this we need to consider the set of feasible contracts rather than just two, and we need to be more specific about the process by which deviant play occurs.

3. A taxonomy of contracts

Suppose contracts differ in their distributional shares and also in the level of total surplus (sum of payoffs) they yield. Some contracts are, in this sense, more efficient than others. This might occur if the use of a particular technology required a distinct set of property rights, which in turn supported a particular equilibrium contract. An example of this technology-institutions mapping is the relationship between the production of sugar and precious metals extraction in the early history of the New World and the exclusive and coercive institutions that prevailed there Sokoloff and Engerman (2000). Another is the rise of agriculture and the emergence of individual property rights

ten or so millennia ago Bowles and Choi (2002).

Analysis of the 2x2 contract game will be facilitated if we write a_{11}=1, b_{11}=1 and $a_{00} + b_{00} = \rho$, so $\rho/2$ is a measure of the relative efficiency of the {0,0} convention; when ρ takes the value of 2, the two conventions produce the same the joint surplus. Further let the A player's share of joint surplus in the B-favoring {0,0} equilibria be $\sigma<½$, with $(1-\sigma)$ the share of gained by B. The payoffs are shown in figure 4.

The contract space shown in figure 5 depicts a set of Alternative contracts defining convention E_0. Point *S'* is the Benchmark contract (with $\rho = 2$ and $\sigma=½$). Thus if the two possible contracts are represented by points *S'* and *x*, both groups will prefer the Alternative contract because both $\sigma\rho$ and $(1-\sigma)\rho$ exceed 1 under its terms. Contracts above AS' are Pareto-superior to the benchmark. (Ignore the locus S'S for the moment.)

Conflict of interest between the two groups is confined to the contracts lying below *AS* and above *BS*. This does not ensure that the *S'* would be eclipsed by an alternative contract like x. The reason is that while x is Pareto superior to the *S'*, adherence to *S'* is a mutual best response and so will only be dislodged by non-best response actions such as the strikes or lockouts mentioned above. Our intuition, however, says that Pareto inferior conventions must be at a disadvantage in a stochastic environment.

If non-best response actions (also called "idiosyncratic") occurred by chance, each person best responding with probability $(1-\varepsilon)$ and adopting the other strategy with probability ε, we can use the results of stochastic evolutionary game theory to make some strong predictions. A striking theorem due to Peyton Young (1998) demonstrates that populations evolving according to this dynamic will spend most of their time at conventions that are not only efficient but also egalitarian.

Under some innocuous restrictions on the updating process, Young's "Contract Theorem" shows that the most persistent convention (called the stochastically stable state) is the one which maximizes the relative payoffs of the group with the lowest relative payoff.

The reason is that conventions that are egalitarian in this sense have larger basins of attraction than more unequal conventions. For this reason, the chance bunching of a sufficiently large number deviant players tipping the population into the basin of attraction of such a convention is likely to happen sooner than for a convention with a smaller basin of attraction. Thus such conventions are *accessible*. It is also true that such conventions are also *robust*, in that they are unlikely to be unraveled by non best response play, because having a large basin of attraction means that it takes a large fraction of one population or the other to unravel the convention. The combination of accessibility and robustness is what determines the persistence of a convention. If we consider an Alternative contract

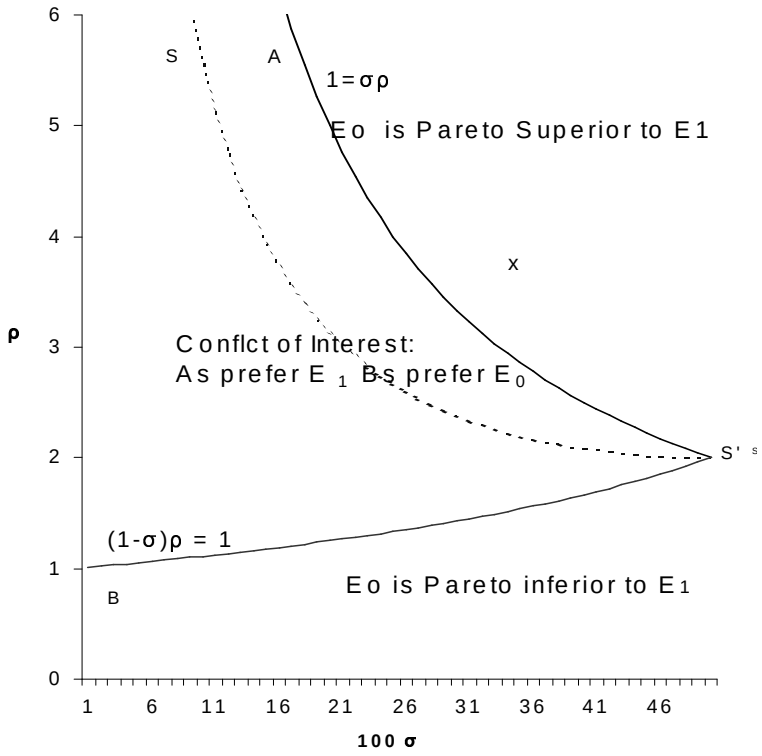

Figure 5 Contrasting contracts. Each point represents the efficiency and distributional share of the Alternative contract supporting the equilibrium E_o. Contracts above *S'A* are Pareto superior to the Benchmark contract with $\rho = 2$ and $\sigma = ½$. Contracts below *S'B* are Pareto-inferior to the Benchmark contract.

in Figure 5 and compare it with the Benchmark {1,1} contract, the two basins of attraction will be of the same size if $a_{00}b_{00} = a_{11}b_{11}$, which using the payoffs in Figure 4 requires that

$$\sigma(1-\sigma)\rho^2 = 1$$

If the alternative convention is such that $a_{00}b_{00} > a_{11}b_{11}$ (meaning $\sigma(1-\sigma)\rho^2 > 1$) then the alternative will be the more persistent of the two conventions. It is clear from this condition that both relative efficiency and equality of shares contribute to stochastic stability of a convention (the term $\sigma(1-\sigma)$ is maximized for $\sigma = ½$.).

Figure 5 illustrates a novel kind of efficiency-equality tradeoff. It depicts the relationship between efficiency and equality as determinants of persistence: *SS'* is the locus of combinations of ρ and σ such that $\sigma(1-\sigma)\rho^2 = 1$. Thus *SS'* is the locus of Alternative contracts such that both conventions are equally persistent. Alternative contracts above *SS'* are more persistent when the other convention is based on the Benchmark contract. For Alternative contracts below *SS'* the Benchmark contract is more persistent. Calculations using the model just described confirm that more efficient and more equal conventions are likely to be more persistent. We can explore the long run behavior of the system by assuming that {0,0} is the status quo and then calculating the expected waiting time (number of periods) before a the idiosyncratic chance actions of either the A's or the B's induce a tipping event, propelling the population from {0,0} to {1,1}. Using the same method we then calculate the expected waiting time for a transition back to {0,0}. The fraction of the "round trip" – that is the waiting time from {0,0} to {1,1} and back – that the population spends at each equilibrium can then be calculated.

Figure 6 gives the results of this calculation where the two sub-populations each have 12 members and for various values of σ and ρ. The height of each bar gives the expected fraction of time the population will spend at the convention in question if the Benchmark contract is {1,1} as before. Where E_0 is identical to E_1 (ρ = 2 and σ = ½ indicated by the dark bar at these coordinates) the population spends half of its time at each convention as one would expect. One can see a band

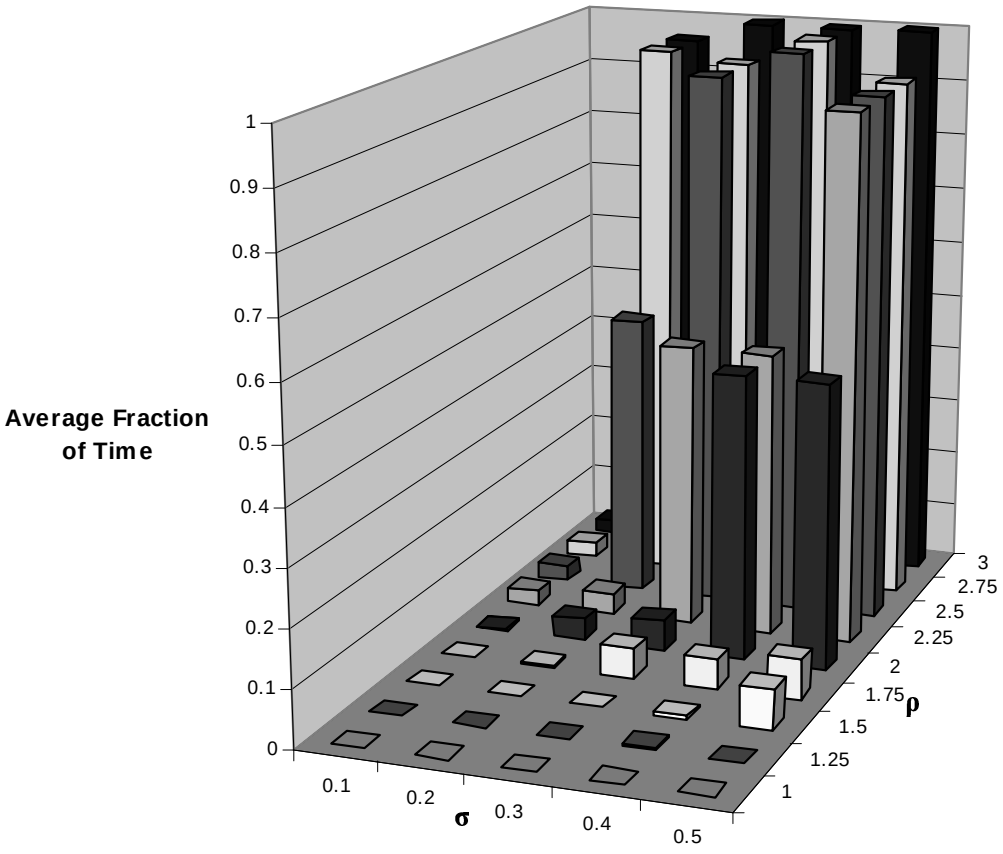

Figure 6 Efficient and equal conventions are persistent when deviant behavior is unintentional. Note: the Benchmark convention is E_1 for which ρ = 2 and σ = ½.

of conventions (similar to the locus *SS'* in Figure 5) which like (ρ = 2 and σ = ½) generate equal

average waiting times (for example, $\rho = 2.5$ and $\sigma = 0.2$ generates this result, as does $\rho = 2.25$ and $\sigma = 0.3$). The population will spend virtually all of the time at conventions more efficient or more equal than these and virtually none of the time at conventions less efficient or less equal.

The reason that more equal conventions are favored in this framework is the following. Consider an Alternative contract with $\rho = 2$ and $\sigma < ½$. An increase in the distributional share of the A's in the Alternative contract has two effects. First, it lowers α^* and thus it requires fewer instances of idiosyncratic play by the A's to disrupt the Alternative Contract, inducing a movement to the Benchmark (which they prefer). The reason is that when the Alternative is less unequal, it takes fewer idiosyncratic A's to induce the B's to switch to the Benchmark. The second effect of an increase in σ is to raise β^* thus reducing the minimal fraction of non best responding B's (namely $(1-\beta^*)$) required to induce the A's to abandon their preferred Benchmark contract in favor of the Alternative. The two effects of a more equal Alternative contract work in opposite directions, the first leading to a shorter waiting time for a transition from the Benchmark to the Alternative, and the second leading to a shorter waiting time for the reverse transition. But for $\sigma<½$ the second effect is larger, so the population will spend more time at the Alternative, the more equal it is.

It is easy to see why efficient conventions would be favored in this setup. For at least one group, offering the efficient contract must be risk dominant in the standard sense that if one believes that the other will offer the two contracts with equal probability, then the best response is to offer the more efficient one. Inefficient conventions are not accessible because it takes a large amount of non-best-response play to induce best responders to shift from an efficient to an inefficient convention. Note that this is not because best responders anticipate the consequences of their switching for the population level dynamics. Rather, their response is purely individual and based

on past (not anticipated future) population states; no individual is attempting to implement the more efficient convention. Inefficient conventions are not persistent for analogous reasons.

Less transparent is the result that highly unequal conventions are not good candidates for persistence. This is a consequence of the fact that they are easily unraveled, because as Young (1998):137 puts it: "it does not take many stochastic shocks to create an environment in which members of the dissatisfied group prefer to try something different." Note that in this example it is the idiosyncratic play of the *privileged* group that unravels the unequal convention, that is, the convention from which they benefit disproportionately. Note also, that the more unequal the convention is, the more easily is it unraveled by this means.

If deviant behavior is indeed just accidental, then this account is correct. But the plausibility of the illustrations given above – the "strike" and the "lockout" – rested on the idea that deviant behaviors are undertaken deliberately and not by mistake. Indeed it typically is the deviant behavior of the disadvantaged group that induces institutional transitions. It was not a fortuitous piling up of unlikely accidents on the part of Communist Party officials that doomed Communism, but rather a combination of chance events and the deliberate and coordinated actions taken by those seeking to live under other institutions. .

4. Collective action

We therefore need to study how the distributional share of a contract (σ) might affect the vulnerability to the collective action of those seeking to improve their share. When intentional non-best-response play is introduced in the form of collective action by those trying to displace the status quo convention, the dynamic of institutional innovation is substantially altered. It is no longer generally the case that the persistent states are egalitarian and efficient. In particular, if the rich are

few and the poor many, unequal and inefficient institutions can be very persistent. The reason is that when non-best-response play is intentional there is just one way (rather than two) that a convention can be overturned (by the actions of those who would benefit more at the other convention), and the larger numbers of the poor militate against a sufficient fraction of them adopting a non best response to displace the equilibrium under which they do poorly.

The collective action approach requires some modifications in the above model. First the players must be assumed to recognize the possibility of transiting to a new institutional setup, and have the ability to anticipate the consequences of their actions on the actions of others. Thus rather than restricting individuals to backward-looking updating, I now introduce a limited capacity to look forward. By *collective action*, I mean the intentional joint action towards common ends by members of a large group of people who do not have the capacity to commit to binding agreements prior to acting (that is, they act non-cooperatively). Examples are strikes, ethnic violence, insurrections, demonstrations, and boycotts.[5]

To clarify the underlying processes, I will first analyze a degenerate case in which individuals participate in a non-best-response collective action when it is in their individual interest that the action take place. Suppose that there is a probability ε that each person is "called to a meeting" at which those attending consider undertaking a non-best-response action. For example, assume the B-

[5] The proviso that play is non-cooperative excludes the degenerate case (with which I begin for purposes of illustration) of groups whose structure allows the assignment of obligatory actions to each of its members. While most successful collective actions include a wide range of selective incentives and sanctions to deter free riding, few if any groups have the capacity to simply mandate group-beneficial behaviors by individual members.

favorable convention {0,0} obtains and some fraction of B's (resulting from the "call") are considering switching to offer a 1-contract instead. But they cannot benefit from switching because they prefer the status quo convention, and destabilizing it -- should sufficiently many of the other possible B-innovators also switch -- could propel them to the alternate convention under which they would be worse off. These potentially idiosyncratic players would thus decline the opportunity to innovate.[6]

By contrast, imagine that the a group of A's were randomly called for deliberation of the merits of a switch away from the governing convention {0,0}, and suppose that should they all adopt a non-best-response, this will be common knowledge. Each then might reason as follows. If they are sufficiently numerous and if all of them switched, the best response for the B's would be to switch as well. Knowing this, should they all switch, they would anticipate the B's response and so would persist in offering 1-contracts in the next period. As a result, the A-unfavorable convention {0,0} would be displaced.

Suppose there are n members of the A population (previously normalized to unity). Because if fewer than $n\alpha^*$ A's were called, there could be no benefit to collective action even if it were uniformly successful, let us analyze the case for which the number called, η, exceeds this critical level, that is $\eta \geq n\alpha^*$. To lend some concreteness to the case let us say that switching means to engage

[6] Favored groups, like the B's in convention {0,0} may deploy informal or governmental sanctions or to minimize idiosyncratic play of their own members. Examples include the shunning and more severe sanctions imposed on whites offering favorable contracts to non whites in racially stratified societies such as apartheid South Africa and the U.S. South prior to the civil rights movement.

with other A's in a strike, refusing to accept any outcome less than a_{11} (all this means is to offer a 1-contract, so the strategy set is unchanged). We can then calculate the expected waiting time before sufficiently many A's are "called" so that a tipping event (a strike) occurs. Unlike figure 6 we do not include the possibility that the B's idiosyncratic actions would induce a "tip" because when sufficiently many of them are "called" they recognized that they would not benefit from disrupting the status quo convention. Then, supposing that the population has been tipped to the {1,1} convention by the A's strike, we calculate the expected waiting time before sufficiently many B's are "called" to allow a tipping event back to the B-favorable convention {0,0}. As in figure 6 the fraction of the "round trip" – the waiting time from {0,0} to {1,1} and back – that the population spends at each equilibrium is then calculated

The height of the columns in Figure 7 gives the fraction of time spent at each of a number of Alternative conventions when the benchmark is as before {1,1} and the unequal Alternative conventions are characterized by differing levels of efficiency. The figure shows the effect of assuming sub-populations of different size (retaining the degenerate model of collective action) for an alternative contract with $\sigma = 0.3$ and with the ρ values as shown. By contrast to the equal sub-population size case depicted in Figure 6, when population sizes differ unequal and quite inefficient conventions may be highly persistent. For example, in the equal population size case a convention with $\sigma = 0.3$ needed a ρ of 2.25 to be equally persistent to E_1; but if the A's number 18 and the B's 6, the two conventions are equally persistent when the unequal convention is much *less* efficient than the benchmark, that is $\rho = 1.25$ Where there are 21 A's (and 3 B's) the population will spend most of the time in the unequal convention even if its level of efficiency is half that of the equal convention. Note that the level of inequality measured by the average income of B's relative to A's

is n(1-σ)/σ(24-n), each B interacting with more A's as their relative share of the population increases. Thus at the convention E_0 if σ = 0.3 and the A's and B's are equally numerous, the B's have an income 2.33 times the A's but when there are 21 A's and 3 B's, the ratio is 16.33. Thus highly unequal distribution of income may result from unequal sub-population sizes, and may be persistent because of the unequal sub-population sizes.

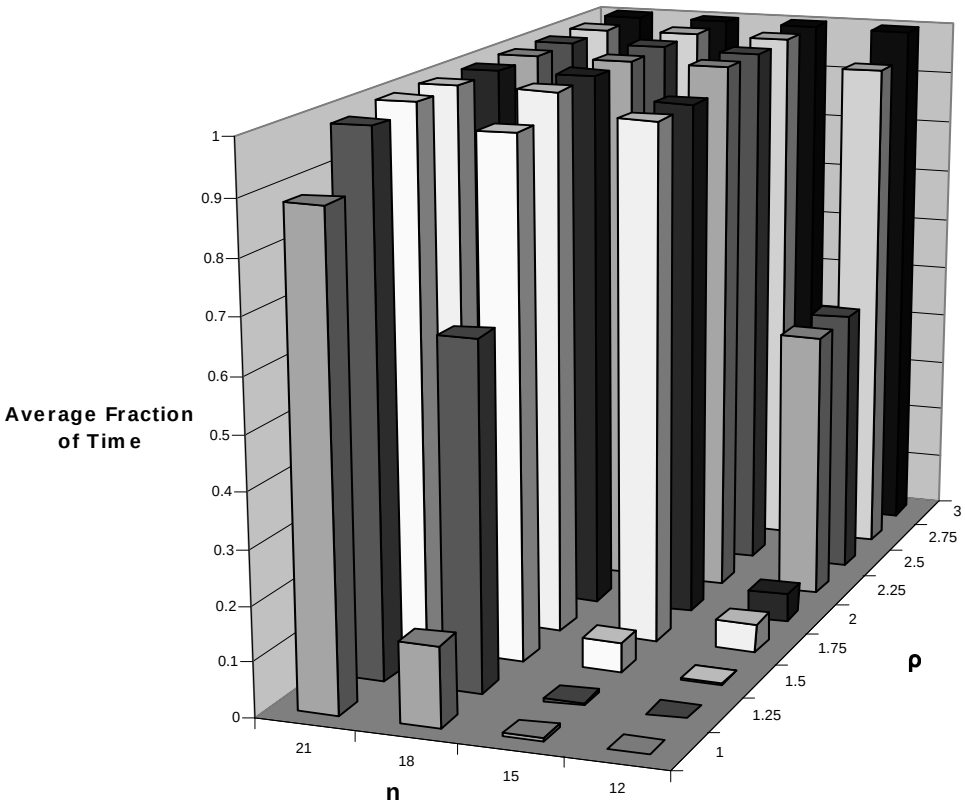

Figure 7 Unequal conventions persist when deviant behavior is intentional and the poor outnumber the rich. Note: total population is 24; the Benchmark convention is E_1 (σ = ½, ρ = 2). E_0 is characterized by the values of ρ indicated and σ = 0.3.

The evolutionary success of unequal and inefficient conventions benefitting the smaller of the two

classes is readily explained. As long as the rate of idiosyncratic play is less than the critical fraction of the population required to induce a transition (which I assume), smaller groups will more frequently experience "tipping opportunities" when the realized fraction of the population who are "called" by chance exceeds the expected fraction (ε itself). The theory of sampling error tells us that the class whose numbers are smaller will generate more "tipping" possibilities. Small size does not facilitate collective action if more than the critical number are "called".

5. Persistent inequality

So far I have abstracted from the problem of collective action by assuming that whenever a sufficient fraction of a sub-population is "called" they will adopt a non-best-response if they (and their group) would benefit if all of those called adopted the non-best-response. But this was just a pedagogical device permitting a sequential presentation of the main aspects of a unified causal mechanism. Thus, what is needed is a model of the coordination problem posed by collective action, nested in the larger population game representing institutional evolution. Taking account of the intentional nature of collective action will provide an explanation of why highly unequal conventions may be vulnerable to unraveling, while moderately unequal conventions may be highly persistent, even if they are not particularly efficient.

Because collective actions generically take the form of *n*-person public goods games in which the dominant strategy is non-participation if preferences are wholly self-regarding, the extended model must address incentives for each to free ride when other act in pursuit of commonly shared objectives. A second desideratum is that the model should reflect the fact that opportunities for collective action often arise by chance, or at least in ways too complex to tractably model, examples being economic depressions, wars, price shocks, booms, and natural disasters. Finally, unlike

idiosyncratic play, participation in collective action is not only intentional (rather than accidental) but is also conditional on one's beliefs about the likelihood and consequences of a substantial number of one's kind changing behaviors. For this reason, facts about global rather than simply local payoff (that is, payoffs both in the present convention and in the alternative, rather than those in the neighborhood of the current population state alone) may have a bearing on the outcomes.[7]

Engaging in this collective activity yields in-process benefits of two types. First, irrespective of the consequences of the action, conformism (or punishment of non-conformists) may impose a cost on those not adopting the most common action. So, let *c* be the cost of being a sole non-conformist, and the conformism costs to those striking being (1-*s*)*c* where s is the fraction of those "called" who strike. The costs to the non strikers is *sc*. Further, there are net benefits or costs associated with the action that may be independent of the numbers participating, including both the time, resources and possibly risk of harm associated with the collective action as well as the positive value of participating, or whatWood (2003) terms the "pleasure of agency".[8]

It is reasonable to suppose that these subjective benefits depend on the magnitude of the gains to be had if the action is successful, not primarily because these gains are a likely consequence of

---

[7] This means that individuals are forward looking to the extent that they can anticipate the consequences of successful collective action.

[8] Compelling evidence from the histories of collective action (e.g. Moore (1978), Wood (2003)) anthropology ((Boehm (1993), Knauft (1991)), experimental economics (Fehr and Gaechter (2002)), and agent-based models of cultural and genetic evolution (Bowles and Gintis (2004), Boyd, Gintis, Bowles, and Richerson (2003)) suggests that individuals knowingly engage in costly actions to punish violations of norms, even when these actions cannot otherwise benefit the individual.

one's individual participation (which is very unlikely in large groups) but because the magnitude of the gains to be had is plausibly related to the strength of the norms motivating the action. The pleasure of participating in a collective action that would if successful transform the conditions of one's class from squalor to abundance is likely to be greater than the pleasure of acting for wage increase a few cents more an hour.

So let the net subjective benefits for an A engaging in a collective action to displace convention {0,0} be

$$\delta=\delta(a_{11}-a_{00})$$

where δ is a positive constant, reflecting the fact that joining a collective action in pursuit of an institutional change from which one and one's peers will not benefit confers no benefits. If the strike fails (because too few participate in it) the status quo convention will persist, and all A's will get a_{00} in subsequent periods independently of whether they participated in the strike or not. Likewise if the strike succeeds all As will get a_{11} subsequent periods, irrespective of their actions this period. Thus the relevant comparison is between the single period net benefits to striking (insisting on contract 1, refusing contract 0) or abstaining are:

$$u_1 = \delta(a_{11}-a_{00}) -(1-s)c$$

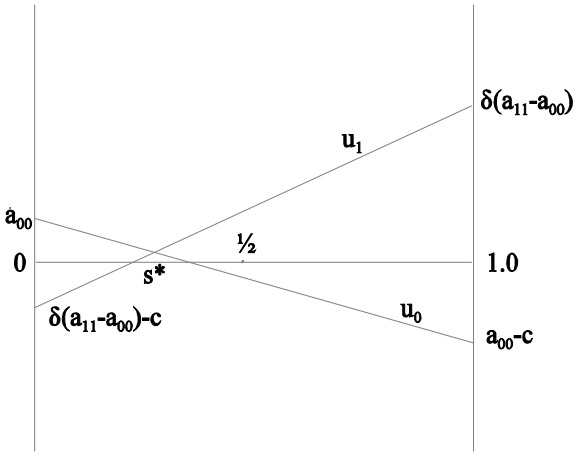

Figure 8 The collective action problem. Note if $s^*<½$ the risk dominant equilibrium is universal participation in the non-best-response action.

$u_0 = a_{00} - sc$

These payoff functions are illustrated in Figure10, from which it is clear if those involved believe that at least s* of their fellows will join in, then strikers' expected payoffs will exceed those of non-participants, and hence all will elect to strike The critical value, s* equates u_0 and u_1:

$s^* = ½ - [\delta(a_{11} - a_{oo}) - a_{oo}]/2c$

How might A's beliefs be formed? The simplest supposition consistent with the above model is that having no information about what the others will do, each believes that the likelihood of each of the others participating is ½, so the expected fraction participating is ½, and all will participate if s^* less than one-half.[9] In a game of this type, the strategy that maximizes expected payoffs when one attributes equal probability to the others' choice of a strategy is termed the risk dominant strategy, and the result of all following a risk dominant strategy is termed the risk dominant equilibrium.

Thus unanimous participation (of those "called") will occur if striking is the risk dominant equilibrium of the collective action game, requiring that the numerator of the bracketed term on the right hand side of (9) be positive, or that the "pleasure of agency" outweighs the loss of a single period's income. (Note that while inferior payoffs in the status quo convention (a_{11}-a_{00} > 0) is a

[9] The choice of ½ is conventional but arbitrary; individuals may have prior beliefs of the fraction likely to participate based on previous similar situation and the like. If individuals then apply their reasoning to each of the others (each, supposing that half will participate, will also participate), they would then correctly predict that s=1; but while this second round of induction may determine whether the individual expects the collective action to be successful in displacing the convention, is not relevant to the individual's behavior, as the relative payoffs of participating or not are independent of the success of the action.

necessary condition for participation, it is not sufficient, as it does not insure that $\delta(a_{11}-a_{00})-a_{00} > 0$.)

The properties of the dynamical system are substantially altered by modeling idiosyncratic play as intentional collective action. Notice that if $\delta(a_{11}-a_{00})-a_{00} < 0$ collective action will not take place (irrespective of the numbers of randomly drawn potential innovators), so the A-unfavorable convention {0,0} if ever attained will persist forever (it is an absorbing state.) Thus in the dynamical system with collective action as the form of non-best-response, institutional lock-ins are possible, with initial conditions determining which of the two conventions will emerge, and then persist forever. To see that this must be the case for a finite "pleasure of agency" parameter δ, consider an unequal convention with $a_{11}-a_{00} = \Delta$ letting Δ become arbitrarily small must make $\delta(a_{11}-a_{00})-a_{00} < 0$ so collective action by A's will not occur and E_0, should it ever occur, will persist forever. Thus there must exist a set of conventions, less equal than E_1 and no more efficient, that will persist forever. These conventions are examples of inequality resulting from institutional lock-in.

How are we to interpret these persistent unequal states? Over relevant time scales, the parameters of the model are likely to change due to cultural and political changes affecting δ or technical or other changes affecting the payoffs to the relevant contracts. Suppose some unequal Alternative contract defines the status quo convention (E_0), and it represents an absorbing state. If technical change made the {1,1} contract progressively more efficient by comparison to {0,0}, then $\delta(a_{11}-a_{00})$ would eventually exceed a_{00}. As a result, the conditions for collective action would obtain, and a transition from E_0 to E_1 would eventually take place. Transitions in the reverse direction would become more unlikely over time as the increase in a_{11} raises the minimum number of non-best-responding B's required to unravel E_1. Thus, the institutional demands of new technologies may account for the emergence of new contractual conventions. A cultural change enhancing the pleasure

of agency, δ – a role played by liberation theology in some parts of Latin America, by the spread of democratic ideology in South Africa and the former Communist countries - would have the same effect. This is very roughly Marx's account which sees history as a progressive succession of "modes of production," each contributing to "the development of the forces of production" for a period, then becoming a "fetter" on further technological advance and being replaced through the collective action of the class which would benefit by a shift to a new convention more consistent with the new technologies.

6. Conclusion

The above reasoning suggests that unequal institutions may persist over long periods due to the nature of these arrangements as self-enforcing conventions and the difficulty faced by the poor in coordinating the types of collective action necessary to "tip" a population from an unequal to a more equal set of institutions. Very unequal institutions would be quite vulnerable to unraveling if the actions inducing the change were the idiosyncratic non-best response actions of members of the class that benefits from these institutions. But this process obviously does not capture the dynamics of most historical transitions. By contrast, in the more empirically relevant case institutional tipping takes place through the deliberate refusal of those disadvantaged by the status quo convention to abide by its terms. In this case highly unequal conventions are difficult to dislodge: the more unequal the status quo is, the greater the number of deviant actions of the least well off are required to tip the population to a more egalitarian alternative. Taking account of the intentional nature of deviance, and the fact that it takes the form of collective action introduces the central question of motivation. The main result is thus that moderate levels of inequality may be insufficient to motivate collective action by *any of the poor*, while conventions characterized by extreme levels of inequality can only

be displaced through collective actions endorsed by *very large fractions of the poor*.

Institutions differ in evolutionarily relevant ways not captured in this model by measures of efficiency, distributional shares, and relative group size, of course. Some institutions may facilitate collective action of the disadvantaged, while others make it more difficult to coordinate the actions of the poor. Marx, and many since, have believed that the social conditions of industrial capitalism constituted a schoolhouse of revolution, by contrast with earlier institutions of sharecropping, tax farming in societies of independent peasants, and slavery, for example. By contrast, Moore (1966), Mao (1953), Markoff (1996) Scott (1976), Wood (2003) and others, with perhaps greater accuracy, have seen patron-client relationships in agrarian societies and highly unequal systems of land holding as especially vulnerable to revolutionary overturns. These extensions of the basic model may be represented in the differing net benefits of collective action, δ, subscripted by the conventions to which they apply.

Whether the persistence of institutions that implement high levels of inequality can be explained by such a model is of course an open question, one awaiting the necessary theory-building and empirical investigation. A plausible alternative explanation of institutional evolution makes success in intergroup competition – victory in warfare, access to favorable environments, heightened rates of population increase and the like -- a key arbiter of the rise, diffusion and demise of institutions. These processes were implicit in Talcott Parson's description of evolutionary universals, and explicit in the related thinking of Frederich Hayek (1988). Models along these lines suggest that some egalitarian institutions -- such as food sharing among non-relatives or monogamy -- may contribute to group success and thus to their own persistence (Bowles, Choi, and Hopfensitz (2003)). More adequate models would encompass both the within-group dynamics modeled here and

these and other forms of between group competition. (An example is provided in Bowles and Choi (2002).)

Works cited

Banerjee, Abhijit and Lakshmi Iyer. 2002. "History, Institutions and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." MIT Working Paper 02-27.

Boehm, Christopher. 1993. "Egalitarian Behavior and Reverse Dominance Hierarchy." *Current Anthropology*, 34:3, pp. 227-54.

Bowles, Samuel. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. Princeton: Princeton University Press.

Bowles, Samuel and Jung-Kyoo Choi. 2002. "The First Property Rights Revolution." *Santa Fe Institute Working Paper 02-11-061*.

Bowles, Samuel, Jung-Kyoo Choi, and Astrid Hopfensitz. 2003. "The coevolution of individual behaviors and group level institutions." *Journal of Theoretical Biology*, 223:2, pp. 135-47.

Bowles, Samuel, Steven Durlauf, and Karla Hoff eds. 2004. *Poverty Traps*: Russell Sage Foundaiton.

Bowles, Samuel and Herbert Gintis. 2004. "The evolution of strong reciprocity: cooperation in a

heterogeneous population." *Theoretical Population Biology*, forthcoming.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter Richerson. 2003. "The evolution of altruistic punishment." *Proceedings of the National Academy of Science (USA)*, 20, pp. 123-43.

Cortes, Hernan, translated and edited by Anthony Pagden. 1986. *Letters From Mexico*. New Haven: Yale University Press.

Eldredge, Niles and S.J. Gould. 1972. "Punctuated Equilibria: an Alternative to Phyletic Gradualism," in *Models in Paleobiology*. T.J.M Schopf and J.M. Thomas eds. San Francisco: Freeman, Cooper, pp. 82-115.

Fehr, Ernst and Simon Gaechter. 2002. "Altruistic Punishment in Humans." *Nature*, 415, pp. 137-40.

Foster, Dean and H. Peyton Young. 1990. "Stochastic Evolutionary Game Dynamics." *Theoretical Population Biology*, 38, pp. 219-32.

Hayek, F. A. 1988. *The Fatal Conceit: The Errors of Socialism*. Chicago: University of Chicago Press.

Knauft, Bruce M. 1991. "Violence and Sociality in Human Evolution." *Current Anthropology*, 32:4,

pp. 391-428.

Mao, Zedong. 1953. *Report of an investigation into the peasant movement in Hunan*. Peking: Foreign Languages Presss.

Markoff, John. 1996. *The Abolition of Feudalism: Peasants, Lords and Legislators in the French Revolution*. University Park: Pennsylvania State University Press.

Moore, Barrington Jr. 1966. *The Social Bases of Dictatorship and Democracy*.

Moore, Barrington Jr. 1978. *Injustice: The Social Bases of Obedience and Revolt*. White Plains: M.E.Sharpe.

Parsons, Talcott. 1964. "Evolutionary Universals in Society." *American Sociological Review*, 29:3, pp. 339-57.

Scott, James C. 1976. *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. New Haven: Yale University Press.

Sokoloff, K. and S. Engerman. 2000. "Institutions, Factor Endowments, and Paths of Development in the New World." *Journal of Economic Perspectives*, 14:3, pp. 217-32.

Wood, Elisabeth. 2003. *Insurgent Collective Action and Civil War In El Salvador*. Cambridge: Cambridge University Press.

Wright, Sewall. 1931. "Evolution in Mendelian Populations." *Genetics*, 16, pp. 97-159.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.