University of Illinois College of Law
Law and Economics Working Papers

Year 2005 *Paper* 23

Conformity to Inegalitarian Conventions and Norms: The Contribution of Coordination and Esteem

Richard H. McAdams*

*University of Illinois College of Law, rmcadams@uchicago.edu

Conformity to Inegalitarian Conventions and Norms: The Contribution of Coordination and Esteem

Richard H. McAdams

Abstract

In this contribution to a symposium on "Conformism," I comment on two of the many mechanisms producing conformity: coordination and esteem. First, I set forth one point about conformity coordination settings – that there can be a strong stability to conventions in which the required behavior varies by the observable physical differences among human beings, such as sex and those that come to be associated with race. In a certain class of important games, observable personal differences work to "break symmetry," which significantly changes the possible outcomes to the game. Second, I explain the claim that human beings desire the esteem of others and then discuss how this simple preference can produce significant conformity. As with coordination, one implication is that esteem-seeking among strangers is likely to make behaviorally relevant the distinctions among individuals that even a stranger will know, i.e., observable physical traits, including sex and race. In both cases – coordination and esteem – I emphasize some inegalitarian (and illiberal) types of conformity.

The Contribution of Coordination and Esteem

forthcoming 88 THE MONIST (2005)(issue on Conformism).

Richard H. McAdams
rmcadams@law.uiuc.edu
University of Illinois College of Law
504 E. Pennsylvania Avenue
Champaign, IL 61820
(217) 333-4385

ABSTRACT

In this contribution to a symposium on "Conformism," I comment on two of the many mechanisms producing conformity: coordination and esteem. First, I set forth one point about conformity coordination settings – that there can be a strong stability to conventions in which the required behavior varies by the observable physical differences among human beings, such as sex and those that come to be associated with race. In a certain class of important games, observable personal differences work to "break symmetry," which significantly changes the possible outcomes to the game. Second, I explain the claim that human beings desire the esteem of others and then discuss how this simple preference can produce significant conformity. As with coordination, one implication is that esteem-seeking among strangers is likely to make behaviorally relevant the distinctions among individuals that even a stranger will know, i.e., observable physical traits, including sex and race. In both cases – coordination and esteem – I emphasize some inegalitarian (and illiberal) types of conformity.

Conformity to Inegalitarian Conventions and Norms: The Contribution of Coordination and Esteem

Conformity is a large topic and its causes are undoubtedly heterogenous. Of the various mechanisms that contribute to conformity, I will comment on two: coordination and esteem. Game theorists have given coordination significant attention. Lewis (1969) first posited that social conventions are, roughly, particular equilibrium outcomes to recurrent coordination problems. Once the equilibrium occurs, it is, by definition, in everyone's interest to conform. Evolutionary game theorists have explored the conditions that make a certain equilibrium likely to emerge and persist when more than one equilibrium is possible (see Sugden 1986, Skyrms 1996, Young 1998). In the first section below, I set forth one point about the nature of conformity in such settings – that there can be a strong stability to conventions in which the required behavior varies by the observable physical differences among human beings, such as sex and those that come to be associated with race. In a certain class of important games, observable personal differences work to "break symmetry," which significantly changes the possible outcomes to the game. My aim is not to provide a particular model of conformity involving sex and race, but to illustrate the usefulness of a particular approach to model-building.

Less studied is a second mechanism of conformity: the desire for "esteem," i.e., the intrinsic (rather than instrumental) concern for one's reputation. In the second section below, I explain the claim that human beings desire the esteem of others. I then discuss how this simple preference can produce significant conformity. Strong patterns of (expressed or unexpressed) approval and disapproval create new incentives for behavior. These attitudinal patterns can themselves create behavioral patterns or, more commonly, provide a new incentive for complying with existing behavioral patterns, which may arise through various processes including coordination. Again, I do not offer a particular model, but rather describe some of the implications of the simple assumption that individuals intrinsically care what others think of them. As with coordination, one implication is that esteem-seeking among strangers is likely to make behaviorally relevant the distinctions among individuals that even a stranger will know, i.e., observable physical traits, including sex and race.

In both cases – coordination and esteem – I emphasize some unattractive (inegalitarian and illiberal) types of conformity. Though the same mechanisms also produce some desirable forms of social order, I seek to balance the optimism of the existing literature.

I. Coordination: "Breaking Symmetry" With Observable Personal Traits

Ullmann-Margalit (1977:134-97) proposed that coordination games model certain conventions of inequality (her term is "norms of partiality"). Recent papers have applied this sort of theory to explain, for example, the sexual division of labor that exists (in varied forms) across cultures (Hadfield 1999) and the practice of female footbinding and genital mutilation. (Mackie 1996). Thus far, however, this line of literature has not generally incorporated the tools of evolutionary game theory. My thesis is that this newer approach shows generally why we should

expect it to be common that conventions arise along the lines of sex and race.

To support my thesis, I draw attention to an important concept in evolutionary game theory – the distinction drawn by Sugden (1986) and Skyrms (1996) between *symmetric* and *asymmetric* games. As Skyrms observes, the symmetry idea is related to Aumann's (1974, 1987) idea of correlated equilibria. Let us begin with that.

Aumann proved that in certain games additional equilibria are made possible if the parties can, prior to their action, mutually observe a random event. Given such observations, the players will sometimes benefit by *correlating* their actions on the outcome of the event. Rather than describe the point technically, I will illustrate it with an example from Brown & Ayres (1994:373-77). They imagine two individuals in the Battle of the Sexes game in Figure 1, which they take to represent a negotiation (where a 50/50 split of the economic surplus is not feasible). This game has two equilibria (OO and BB, with payoffs of 1 or 5) and the two players each prefer reaching either equilibrium to either of the two non-equilibrium outcomes (OB and BO, with payoffs of 0). But the players also have conflicting preferences because each prefers a different equilibrium.

		Player 2	
		O	B
Player 1	O	5,1	0,0
	B	0,0	1,5

***Figure 1*: A Battle of the Sexes Game**

In this situation, Brown & Ayres observe that the parties would want to correlate their actions with a random event, such as a mediator's recommendation based on a coin flip. There are two possibilities: the players could correlate the outcome "heads" with the OO equilibrium and "tails" with BB, or they could do the opposite. In either case, the mutual correlation of actions gives each a 50% chance of getting his best outcome, a 50% chance of getting his second best outcome, and a 0% chance of a non-coordinated outcome. Without correlation, the only symmetric equilibrium is the mixed strategy equilibrium where player 1 plays O 83% of the time, player 2 plays B 83% of the time. The result is that they fail to coordinate 72% of the time, each receiving (0,0), and the expected payoff for each is .83. But if the parties correlate their actions with the coin flip, they never fail to coordinate and the expected payoff for each rises to 3. Thus, the mutually observed random event (or public signal) creates the new strategies of correlating one's action with the event, and the new strategies create new mutually preferred equilibria.

Sugden (1986) focuses on three games other than Battle of the Sexes – Hawk-Dove, Attrition, and Nash Bargaining – that share its key features: multiple equilibria, disagreement over which outcome is best, and agreement over which outcome is worst. Most importantly, though Sugden does not discuss Aumann, he appears to discover a related principle – that the players might coordinate by mutually observing *asymmetries* in their situation that distinguish their roles. One might say they observe a deterministic (rather than random) public signal prior

to taking their action. Consider Figure 2, which presents a Hawk/Dove game. Each player chooses whether to play an aggressive "Hawk" strategy or a submissive "Dove" strategy. The pure strategy Nash equilibria are Hawk/Dove and Dove/Hawk. Each player prefers to be the one to play Hawk against Dove, but they both want to avoid Hawk/Hawk, which is the worst outcome for each.

	Dove	Hawk
Dove	1, 1	0, 2
Hawk	2, 0	-2,-2

Figure 2: A Symmetrical Hawk/Dove Game

One of Sugden's illustrations is the "Crossroads game" where two cars on different roads approach an intersection at the same time. Imagine that the traffic light is broken or that this is an unregulated territory without traffic rules. In this context, Hawk is the strategy of driving through the intersection, Dove is the strategy of stopping, and the Hawk/Hawk outcome is a collision. Another example is the "Firewood game" where two individuals make incompatible claims to the same resource. In a state of nature, Hawk is the strategy of insisting on the firewood, Dove is the strategy of deferring to the other's claim, and the Hawk/Hawk outcome is violence. Though the winner of the fight receives a payoff higher than the loser, the expected payoffs might be equal (here, -2) if there is an approximately equal chance of winning the fight.

Sugden asks what equilibrium will emerge in a large population if the game is repeated. He supposes that in each iteration of the game, two players from the population are selected at random and play against each other once. The particular equilibrium Sugden derives need not detain us. Instead, what is critical is to see his point that the answer will depend on whether the game is *symmetric or asymmetric*. As the game stands in Figure 2, it is in an important sense symmetric. There is nothing in the game that distinguishes the two players and therefore nothing that distinguishes the outcome Dove/Hawk from the outcome Hawk/Dove. All one can say to describe either equilibrium is that one player selects one action and the other player selects the other action. As a result, we can express all the possible strategies in the iterated game merely by selecting a given probability for each action. Thus, a player selects a value for p where Hawk is played with probability p and Dove with probability $(1 - p)$. There are only two pure strategies:

Pure Strategy (1): Play Hawk with certainty ($p = 1$); and

Pure Strategy (2): Play Dove with certainty ($p = 0$).

In equilibrium, each player will select some value p.

By contrast, Sugden says, suppose the payoffs remain fixed but the players notice some asymmetry that distinguishes their roles. *Any* commonly recognized asymmetry in the player's roles means that the players can now choose to play a strategy dependent on which role they occupy. In the Crossroads game, the players might mutually notice that only one driver is on the larger road, is driving the larger car, or is on the right. In his Firewood game, the players might

mutually notice that only one claimant is older than the other, has a name that is alphabetically prior to the other, or is currently in possession of the disputed firewood. These observations formally distinguish the players. Each may perceive that one is "the driver on the left" and the other is the "driver on the right." Or one is the "older player" and the other the "younger player." However trivial the difference appears, it may be sufficient to change the equilibrium of the iterated game.

To illustrate the power of asymmetry (or if one prefers, the power of a public signal based on perceived elements of the game rather than an external random event), I will use the much-discussed and important example of property, adapted from Sugden (1986), Hirshleifer (1987), and Skyrms (1996). Figure 3 shows the same Hawk/Dove matrix as Figure 2, except that the two players are now distinguished as "possessor" and "non-possessor." The labels reflect nothing but the fact that prior to their action the players commonly observe this distinction in their roles – one currently possesses the firewood and the other does not. But this difference is sufficient to change the game. Because the players commonly recognize the asymmetry, they now have an expanded strategy space. In addition to the symmetric strategy set noted above – playing Hawk with probability p and Dove with probability $(1 - p)$ – each player can now also choose from a new set of asymmetric strategies defined by q and r, as follows: *When possessor,* play Hawk with probability q and Dove with probability $(1 - q)$, and *when non-possessor,* play Hawk with probability r and Dove with probability $(1 - r)$. In addition to the two pure strategies noted above, we now have two more:

Pure Strategy (3): Play Hawk with certainty when possessor and Dove with certainty when non-possessor ($q = 1$ and $r = 0$); and

Pure Strategy (4): Play Dove with certainty when possessor and Hawk with certainty when non-possessor ($q = 0$ and $r = 1$).

		"Non-Possessor"	
		Dove	Hawk
"Possessor"	Dove	1, 1	0, 2
	Hawk	2, 0	-2,-2

Figure 3: An Asymmetric Hawk/Dove Game

Most importantly, the new pure strategies create new Nash equilibria that did not exist in the prior game (of Figure 2): where all play Strategy (3) or all play Strategy (4). At these points, no player can gain from unilaterally switching strategies. For example, if everyone else is playing strategy (3), then when you are the possessor you expect the non-possessor to play Dove, and when you are the non-possessor you expect the possessor to play Hawk. Given these expectations, your best reply is to play Hawk when possessor and Dove when non-possessor, which is to say that your best reply is to play Strategy (3) like everyone else. When everyone plays Strategy (3), we have a property-like convention in which everyone behaves as if the possessor of a resource "owns" it. "Property" emerges spontaneously from the interaction of selfish individuals. Of course, the opposite convention – where everyone plays Strategy (4) – is

also possible (though Sugden claims is less probable because of likely payoff differences between possessor and non-possessors).

Skyrms (1996:63-79) observes the connection between Aumann's concept of a correlated equilibrium and the idea of "breaking symmetry" in a game of this sort. A formally random independent event is only one way to provide a signal around which the players correlate their strategy. The game itself may provide public signals of the players' roles that they perceive as approximately random, in which case they can correlate the strategies with these signals.

Skyrms' analysis uses a different methodological tool. Where Aumann and Sugden assume that perfect rationality exists and is common knowledge, Skyrms uses "replicator dynamics," in which it is assumed that the payoff to a given strategy received in round n positively influences the number of players using that strategy in round $n + 1$. The transmission of successful strategies might occur in a variety of ways, including the intuitive possibility that individuals with imperfect reasoning abilities learn by copying the strategies of individuals who, in the prior round, had higher payoffs. Replicator dynamics offer a way of comparing the strength of different Nash equilibria: given random starting points – where all individuals in a population begin with randomly selected strategies – some equilibria are more likely to emerge than others. The starting points that end up in a particular equilibria are said to be within its "basin of attraction." (Though there are other important evolutionary models, such as those with persistent mutation (Young 1998), I will limit my comment to replicator dynamics).

Let's now extend Skryms' analysis to consider symmetry-breaking in a Bargaining game he discusses, inspired by Nash (1950). In this Bargaining game, two players divide some resource. Each must decide simultaneously with the other how much of the resource to claim. If the two claims sum to less than or exactly 100% of the resource, then each receives the amount he claimed (neither gets any unclaimed amount). If the two claims total more than 100%, each receives nothing. In experimental games of this sort, individuals tend to claim exactly one-half. Skyrms uses these results as illustrative of a general tendency to act fairly: splitting gains equally is fair; it recognizes the moral claim of the other.

Skyrms (1994, 1996:1-21) offers an evolutionary account of the origins of this fair behavior. He shows the success of a fair strategy in an iterated version of the Bargaining game. For simplicity, he compares just three strategies: *Greedy*, which claims 2/3 of the resource; *Fair-minded*, which claims ½; and *Modest*, which claims 1/3. Using replicator dynamics, he ran various computer simulations to see how the strategies would do, given a variety of randomly selected starting points. Skyrms found that the Fair-minded strategy "took over" the population in 62% of the simulation runs, but the remaining runs produced a polymorphic outcome of Greedy and Modest strategies. Because the average return in these cases is only 1/3, compared to the return of ½ that Fair-minded receives when everyone plays it, Skyrms refers to these inefficient outcomes as "polymorphic traps."

Skryms then advances a reason for optimism greater than that warranted by the 62% outcome. Although his basic framework is the random pairing of individuals for each iteration of

the Bargaining game, he proposes that there will be some positive correlation in the strategies that interact (because, for example, those employing similar strategies tend to live near one another). If so, then this is big advantage for Fair-minded, which always gains ½ when it is paired with itself and a big disadvantage for Greedy, which always gains 0 when paired with itself. He then shows that with a modest positive correlation, Fair-minded takes over the population no matter what the starting point. If people employing a given strategy are slightly more likely to interact with others like themselves, the polymorphic traps disappear.

Some have criticized Skyrms for considering only the possibility of zero or positive correlation and ignoring the possibility of *negative* correlation. (See D'Arms, Batterman & Gorny 1998). If, for example, those playing the Greedy strategy were *less* likely to meet each other than to meet other strategies, then the payoffs to Greedy would rise (encountering Greedy less often means encountering, among others, Modest more often) and the payoffs to Fair-minded (which is now encountering, among others, Greedy more often) would fall. Why would the correlation be negative? These critics imagine that those playing Greedy would invest resources in gathering information to determine the strategy the others play, and then to avoid encountering Greedy. The result is that the polymorphic traps are a bigger problem than they first appeared to be.

There is, however, a simpler and more fundamental explanation of negative correlation than the investment in information: *asymmetry*. Despite his contribution to the importance of "breaking symmetry" in the iterated Hawk/Dove game, Skryms does not discuss asymmetry in his analysis of the Bargaining game. He justifies this omission by stating (1994: 313) that "it is only in situations where the roles of the players are perceived as symmetric that we have the clear intuition that justice consists in share and share alike." One might disagree with this point by noting that, if the Bargaining game is to provide a fundamental explanation for the human sense of fairness, then one must consider how frequently the game will be perceived by the players as symmetric. Even if Skyrms is right to ignore the issue, however, it is still useful for understanding conformity to consider the consequences of common (but morally arbitrary) asymmetries in this game. As we shall see, where Skyrms labels symmetry in the Hawk/Dove game a "curse" (because it frequently leads to the worst outcome), the curse in the Bargaining game is the breaking of symmetry.

Skryms presents the Bargaining game as symmetric. Nothing distinguishes the roles of the players. Because we are assuming that strangers play against each other, their total number of possible strategies is quite limited: players can only choose x where x defines either the amount of the surplus they will claim in each iteration or the average amount they will claim. (The latter refers to the possibility that a player claims different amounts in different iterations, but settles on an average amount he will claim and the degree of variance around that mean).

Perhaps it seems that nothing in the Bargaining game *could* distinguish the roles of the players. Nothing in the bargaining situation clearly does so. By contrast, recall that in the Firewood game, when contested property is possessed by one player, each may observe the distinction between the possessor and non-possessor. Similarly, in the Crossroads game, when

two drivers approach an intersection, one driver is "on the left" and the other is "on the right." Sometimes, however, the nature of the situation offers no obvious asymmetry. The Bargaining game is an example.[1]

Nonetheless, even if nothing in the strategic *situation* creates an asymmetry, there is no guarantee that the Bargaining game will be symmetric. The reason is that the characteristics *of the players* can create an asymmetry, at least in a substantial number of cases. By player "characteristics," one might refer to the history of how a player has behaved in past iterations, but that possibility is ruled out by the models Skyrms and Sugden use, which assume interactions between strangers. Yet even if individuals don't know each others' histories, if the interaction is face-to-face (as is usually assumed), then they can observe certain physical characteristics of their counterpart. Observable personal traits are an important means of breaking symmetry because they are available in any interaction to which both participants are physically present (and sighted), even in situations that are symmetric and even when the individuals are strangers to each other. Personal trait differences break symmetry when nothing else does. If so, then it is possible in all of the games discussed that a stable equilibrium will emerge that involves strategies based on such traits.

Consider again the Bargaining game. Suppose that before players select their action in a given iteration, they commonly observe whether the other is right- or left-handed (assume this is obvious because a person's favored arm is more muscular). Then the possible strategies are as follows. As before, players can choose from a set of symmetric strategies defined by x – the amount of the surplus one will claim in each iteration or on average. But now they can also select from a set of asymmetric strategies. The new pure strategies are defined by y and z, as follows: When one's counterpart is right-handed, claim y, and when one's counterpart is left-handed, claim z.[2] Thus, the mutually observed asymmetry permits one to play different strategies against right-handed and left-handed players. A right-handed player might, for example, play Fair-minded against right-handed players and Greedy against left-handed players.

The new pure strategies again create new pure strategy equilibria. In the symmetric game it can be an equilibrium for everyone to play the strategy Fair-minded against everyone, or for some percentage to play Modest and the rest to play Greedy. But besides these possibilities, we now also have asymmetric equilibria of this sort: all right-handed players play Fair-minded against right-handed players and Greedy against left-handed players, while all left-handed players play Fair-minded against left-handed players and Modest against right-handed players.

Because asymmetry works like negative correlation, these polymorphic equilibria are disturbingly stable (see Sugden 1990:780). Recall that, in the symmetric version of the game, a player using Greedy will play against another player using Greedy or Fair-minded some percentage of the time, with the result that each receives zero. This imposes a cost to Greedy, which renders it vulnerable to Fair-minded. But in the asymmetric version, in equilibrium, a player can identify which players use which strategies, not by knowing their past history, but merely by recognizing their physical traits. Thus, the right-handed player never uses Greedy against someone playing Greedy or Fair-minded, but only against left-handers who, in

equilibrium, only play Modest against right-handers. Imagine a mutation in which a right-hander attempts to play Fair-minded against left-handers. The mutation will die out because this player receives a payoff of only ½ in such encounters compared to the existing right-handers who receive 2/3. The same is true of a mutation in which a left-hander players Fair-minded against right-handers. That strategy receives a payoff of zero instead of the 1/3 payoff received in such encounters by existing left-handers.

I also hypothesize but will not prove that the "basin of attraction" for these equilibria is large. Although the efficiency of an outcome does not guarantee that it is an equilibrium, or if it is an equilibrium, that it has a large basin of attraction, efficiency can work in favor of both characteristics. In the *symmetric* Bargaining game, efficiency gives a powerful advantage to the Fair-minded equilibria. "Waste" occurs whenever the players claim anything but exactly 100% – a higher claim means each receives zero and a lower claim leaves some potential surplus unclaimed. In the polymorphic equilibria where some play Greedy and some play Modest, there is waste whenever two Greedys meet (and receive nothing) or two Modests meet (and leave 1/3 unclaimed). This waste lowers the expected payoffs of both players and therefore lessens the "pull" of the polymorphic equilibria. But just as there is no waste when everyone plays Fair-minded, there is no waste in the *asymmetric* equilibria I have just described because each interacting pair claims exactly 100%. Thus, the polymorphic equilibria in the asymmetric game will exhibit an attracting power much greater than the polymorphic equilibria in the symmetric game.

Obviously, the left- and right-handed asymmetry is not the one of interest. Everything I have just said about that observable physical difference carries through for more interesting asymmetries, such as those of sex and the cluster of physical differences referred to (along with other characteristics) with the terms "race" and "ethnicity."

To illustrate: In the Bargaining game, while it is an equilibrium for everyone to play Fair-minded, it is *also* an equilibrium for members of race or ethnicity *A* to play Fair-minded against other *A*s and Greedy against members of race or ethnicity *B* while *B* members play Fair-minded against other *B*s and Modest against *A*s. *B* members are thus systematically disadvantaged in bargaining with *A*s; over time, they will wind up with less wealth.

Similarly, in the Battle of the Sexes Game, it is an equilibrium for males to correlate their strategy with some random event when paired against other males, and for females to do the same when paired against other females, but for males to always play the strategy associated with their preferred equilibrium when paired against females, who in turn always play the strategy associated with their less preferred equilibrium when paired against males. (The opposite is also an equilibrium.[3]) The game might reflect, for example, household bargains over domestic work responsibilities or leisure expenditures, where each woman conforms to the convention of letting her mate get his preferred outcome because she otherwise expects to incur the costs of non-coordination. Finally, returning to the iterated Hawk/Dove game, the following is an equilibrium: When males play against females, males play Hawk and females play Dove; and when males play against males and when females play against females, possessors play

Hawk and non-possessors play Dove. In this equilibrium, a woman will cede any resource to a man and therefore wind up without any property (or without anything a man would value). Perhaps her only way to enjoy property will be to use the property of a male relative or mate.

Despite the inequality and unfairness of these conventions, they are stable. Each individual is playing his or her best strategy given what the others are doing. That the equilibrium disadvantages, say, women does not give an individual woman any incentive to deviate. To the contrary, if the man is expected to play Hawk, the woman is best off playing Dove. Women can collectively gain only if they act collectively to change the expectations underlying the convention. If all or most women at the same time start to play Hawk whenever they possess a resource (say as part of a social movement), this would produce a lot of painful Hawk/Hawk outcomes and eventually men would no longer expect them to play Dove in that situation. Although men would, as a group, wish to resist the change, they might eventually begin to play Dove. But short of this sort of risky collective action, no individual woman gains by playing the unexpected strategy, even though the expected strategy is inegalitarian.

The theory here is consistent with many other theories of racial and sexual differences. In a sense, however, it is more fundamental; at least it is more parsimonious. We can begin with nothing but common knowledge that players *notice* certain differences among themselves – skin color, the shape of the eyes and lips, genitalia, etc. If so, this fact alone is then sufficient to produce conventions in which the way one plays the game depends on the personal physical traits of the person one is playing against (any one trait or some combination of traits). It isn't initially necessary to assume any differences in the preferences, abilities, or opportunities of people of different sexes or races. Nor is it necessary to assume any difference in beliefs, save one. It is only necessary that, at some point, the players believe there is a statistical correlation between the race or sex of an individual and the way s/he plays the game. The replicator dynamics take over, and the result is a convention based in part on observable traits.

Nonetheless, the personal-trait-as-asymmetry theory is ultimately partial and benefits greatly from connections to other theories. This is true in two respects. First, an asymmetry will not influence the learning of parties unless it is *salient*. Given bounded rationality and a very large number of potentially relevant asymmetries in most types of strategic settings, human beings are not likely to notice the correlations between all possible asymmetries and behavior. Though the desire to procreate (among other things) may explain why sex categories are salient, other theories are needed to explain why other differences in physical characteristics – the ones taken to define race or ethnicity – are salient. This does not mean that one needs an entire theory of race or ethnicity before the asymmetry theory comes into play. One need only explain why human psychology gives mental attention to observable trait differences, and asymmetry then explains how any such attention can produce a large behavioral regularity based on that trait.

Second, even when some asymmetry is mutually salient, the personal-trait-as-asymmetry theory is neutral about *which* of the multiple equilibria will emerge. In the Bargaining game, for example, there is nothing in the theory that makes it more likely to observe a convention in which males play Greedy and females play Modest than the reverse. The same is true of ethnic

or racial differences. This indeterminancy is similar to the indeterminacy in the theory of the evolution of property in the iterated Hawk/Dove game. Explaining why we arrive at the pro-property convention requires additional analysis – such as the claim that the payoffs to possessors and non-possessors differ in certain ways. Similarly, additional theory is necessary to explain not just why physical traits are important to behavior, but why they have come to play the particular role they have in observed societies. But the personal-trait-as-asymmetry theory is a useful starting point for understanding the ubiquity of conventions that incorporate sex and race.

II. Esteem: The Origin of "Nosy" and Discriminatory Norms

I have referred to the equilibrium behaviors in the evolutionary models as "conventions." In this section, I contrast conventions with "norms." I will not offer anything like a full definition of either term, but only note some advantage to using the terms to refer to behavioral regularities that differ based on whether they are supported in part by normative beliefs (see McAdams 2001). On this view, conventions are behavioral regularities that do *not* require that anyone hold any particular normative belief. As the term is often used (more broadly than Lewis 1969), a convention is the Nash equilibrium (or a certain subset, such as a pure strategy Nash equilibrium) that emerges among individuals when more than one equilibrium is possible. Under rational choice models of equilibrium selection, a player first decides what other players will do, not what they *should* do. The player then determines what reply to the others' strategies will maximize his expected utility, but he need have no beliefs about the normative appropriateness of his own behavior. Under replicator dynamics, a player copies successful players without forming any particular beliefs other than those identifying their strategy and level of success.

These models are useful precisely because they demonstrate how much conformity may exist without complex normative beliefs. By contrast, when normative beliefs support the behavioral regularity, I use the term *norm*. I follow Pettit (1990) and usage in various social sciences by reserving that term for behavioral regularities that – unlike a convention – are supported in part by a set of normative attitudes that approve of conformity to the regularity and/or disapprove of nonconformity. That the attitudes "support" the regularity at least in part means that they contribute causally to its existence, but they need not be a cause of the regularity first coming into existence. The attitudes may only make the regularity more resilient to change.

The most obvious sort of normative attitude is one of obligation – believing one is morally obligated to engage in or refrain from certain behavior. The clearest way this attitude might influence behavior is that it is "internalized" so that a person suffers "guilt" when he acts contrary to his moral beliefs or enjoys "pride" from adhering to them. When internalization is widespread, and produces a pattern of behavior, the result is a norm. The norm-based behavioral regularity is – unlike a mere convention – supported by beliefs backed by guilt and pride.

Norms so often consist of internalized obligations that some theorists treat norms *as* internalized obligations (see Coleman 1990). However, Pettit (1990) observes that internalization is not necessary for a behavioral regularity to be supported, in part, by normative

attitudes. Another mechanism is the pursuit of "esteem." That people desire esteem means that they value their reputations intrinsically as well as instrumentally (see Brennan & Pettit 2003; Cowan 2002; Pettit 1990; McAdams 1997). In other words, people gain utility directly from having others think relatively well of them (in general and/or on some specific dimension) and lose utility directly from having others think relatively badly of them. Where a person who has internalized the norm will experience guilt from violating it regardless of whether anyone else discovers the violation, a person who cares only about esteem will experience an intrinsic loss only if he knows that others have discovered (or believes they have discovered) his violation. In pre-20th century political discourse, this motivation was well accepted (Brennan & Pettit 2003:23-25), and it is consistent with much contemporary social science (McAdams 1997:356). Yet esteem is neglected in modern literature on social norms and other causes of conformity.

If people seek esteem, then under very plausible circumstances, approval and disapproval will influence behavior. Some theorists reject this possibility, claiming that the costs of expressing disapproval would deter anyone from expressing disapproval unless a norm forbidding certain conduct already exists, in which case the desire for esteem does not explain the norm. The costs of expressing disapproval are, among other things, the time taken to do so and the risk of "counter-disapproval." Similarly, there are costs to selectively expressing approval – because the absence of approval is interpreted as expressing disapproval – so absent a pre-existing norm, the argument is that everyone will claim to approve of everything.

The objection assumes, however, that one must *express* disapproval in order for the pattern of disapproval to influence behavior. To the contrary, one may expect or infer disapproval without it being expressed. One mechanism is introspection; a person may expect disapproval for conduct – lying, promise-breaking, assault – that he himself disapproves (Pettit 1990). Another mechanism is an inference made from the behavior of others (McAdams 2000: 350-54). Because most people prefer to behave differently toward those they esteem and those they disesteem, as by socializing with the former more than the latter, it is costly to act otherwise in an effort to conceal one's esteem judgments. As a result, a person motivated by the desire for esteem and fear of disesteem will be able to make inferences about what others around him approve or disapprove based on their conduct. These inferences will sometimes be wrong, but they will likely be positively correlated with the actual approval patterns and, in any event, they create new incentives for behavior.

Esteem-seeking can produce a norm in two ways. First, a behavioral regularity may initially arise as a convention and only later become a norm as people begin to approve of conformity to the convention and/or disapprove of non-conformity. Indeed, Lewis' (1969) definition of a convention, though quite a bit narrower than the meaning I have been using, is highly useful for seeing how conventions become norms. Lewis required not merely a Nash equilibrium when more than one is possible, but also (among other elements) the condition that each player prefers that every other player select their equilibrium action. For example, once we coordinate on an equilibrium in driving by selecting the left or right side of the road, every player not only prefers to play his equilibrium strategy, but also prefers that *others* play *their* equilibrium strategy. As Sugden (1998) points out, under this condition, expectations of how

others *will* behave can easily grow in to normative expectations about how others *should* behave. That another harms me by acting contrary to my expectations is often enough to earn my disapproval. Once that occurs, then the normative attitude will turn the convention into a norm, lowering the payoffs to non-compliance, and making the regularity even more resilient. Later, the norm may become internalized, thereby acquiring even greater stability.[4]

Pettit (1990) and McAdams (1997) observe a second way the esteem motive can produce a norm. Suppose that at time 1 there is no pattern of approval or disapproval for some behavior and no behavioral regularity. At time 2, based on new information, a discussion about the behavior produces a normative consensus. For example, the new consensus is that recycling contributes to public welfare. The consensus raises the benefits of recycling because it is now clear that this behavior will earn esteem. At time 3, the greater benefits from the behavior induce enough people to engage in that behavior that we now have a norm. Because the influence of esteem is (at least partly) relative, once a certain number of people engage in the behavior, those that do not now suffer disapproval, which may encourage additional conformity. Thus, the pattern of approval or disapproval can precede and cause the behavioral regularity.

Brennan & Pettit (2003) give a comprehensive analysis of the "economy of esteem," answering important questions about the nature of the demand for and supply of esteem. Though they do not ignore the possibility that esteem-seeking will lead to undesirable outcomes, it is fair to say that they stress the benefits of esteem. I agree that a concern for esteem supports the institutions of civil society that check certain anti-social behaviors, but here I wish to emphasize some other, unattractive consequences of the pursuit of esteem.

What appears useful about norms is their ability to regulate externalities. Coleman (1990:249-60), for example, claims that norms arise out of a "demand" caused by externalities. For example, suppose that among suburbanites good lawn care produces positive externalities – a gain to one's neighbors who appreciate looking out onto attractive yards (though the point I'm making works equally well for negative externalities). Without esteem, one might expect that the amount of effort invested in lawn care to be too low. An individual will invest in such activity only up to the point where his private costs equal his private benefits, but social welfare would be enhanced by his spending more, up to the point where his costs equal his and all his neighbors' benefits. Let PC be the costs *A* incurs caring for his lawn, which I assume for simplicity are equal to the social costs (SC) of the lawn care. Let PB be his private benefits from lawn care, EB the external benefits to his neighbors, and SB the combination of the two – the full social benefits. The problem is that he will maximize his return by investing to the point where PC = PB, but social welfare is maximized at PC = SB. The shortfall in investment is (SB - PB) = EB.

What happens when we introduce the preference for esteem? Possibly nothing. One cannot just will oneself to think well of someone; esteem judgments are essentially reflexive. Purely as a descriptive manner, however, human psychology frequently supports the egoistic possibility that *B* esteems *A* for engaging in conduct that benefits *B* (and disesteems *A* for conduct that harms *B*). Another, more moralistic possibility is that *B* esteems *A* for sacrificing his

own welfare to promote the welfare of others. If either of these relationships holds for *A*'s neighbors, then they will esteem *A* more the more he invests in his lawn care. Esteem then creates an incentive to behave in a way that creates positive externalities (or, conversely, to avoid creating negative externalities).

Call this new esteem incentive *E*. We can now expect the esteem incentive to induce *A* to invest additional effort in lawn care, to the amount of $PB + E$. All of this sounds promising, except that *there is no particular relationship between the strength of esteem incentives in a given context and the additional amount of the behavior (or restraint) that is optimal*. For esteem to "solve" the problem, by inducing optimal investment in lawn care, it would have to be the case that the new esteem benefit exactly equaled the shortfall in his investment, i.e., that $E = EB$. But that seems exceedingly unlikely. The relationship would occur if those granting esteem could manage to grant exactly the level of esteem that induces others to engage in the optimal behavior, but there is no reason to think that reflexive esteem judgments work like this. Even if they did, the computational problem would be enormous: one must determine the "right amount" of esteem for each neighbor to grant, given that *A* is likely to place different values on esteem received from different neighbors.

One might think that esteem nevertheless always improves upon the existing situation by increasing the behavior (or restraint) that is otherwise too scarce, even if it does not increase it to the optimal level. But there are two problems. Coleman (1990:273-78) notes the first, which he terms zealousness. The process creating a norm may lead to an *excess* of the targeted behavior, by overshooting the optimum. Esteem competition is useful for understanding this phenomenon. Because esteem is relative, the amount of esteem "earned" by a given level of behavior depends on how that level compares to what the average person in the population does. If so, then the first person who supplies greater-than-average effort to their lawn not only earns esteem but draws esteem away from everyone else. Some of those who suffer this loss may respond by increasing their investment in lawn care, thus drawing esteem away from those who do not respond. The effect of esteem competition is complex: it might end with the supply of the behavior still lower than is socially optimal, but it might increase to a point beyond what is socially optimal. Indeed, esteem competition might increase the level of behavior so far beyond the optimum that society is worse off than it was with no norm and suboptimal investment.

There is a second problem (see McAdams 1997:412-19). When behavior in the absence of esteem considerations would be desirable, any esteem-based conformity is undesirable. Yet the presence of an externality does not necessarily produce undesirable behavior. With certain assumptions, I could make the point with the lawn care example, but let us turn instead to a more important possibility. Sometimes people have preferences regarding *another's* consumption. Thus, *B* might have a preference that his neighbor or co-worker *A* consume the same beer that *B* consumes or is a "fan" of the same local football team. More disturbing examples are that *B* prefers that *A* worship at the same church or date someone of their (*A* and *B*'s) own race. In these examples, *A*'s behavior creates externalities – the costs and benefits incurred by those whom have preferences about *A*'s consumption decisions. But there may be no persuasive normative argument that *A* should conform to the preferences of others. Even a social welfare analysis may

conclude (at least for discontinuous consumption decisions such as the selection of a church or mate) that *A*'s preference for his own consumption behavior is so much stronger than the combined preferences others have for *A*'s consumption behavior, that the social optimum occurs where *A* satisfies his preference of this sort without regard to their preferences.

It is in exactly this case that esteem-based *nosy* norms can arise. As before, *B* will frequently disesteem *A* for engaging in behavior that lower *B*'s utility, which includes *A's* making the consumption choices that *B* disprefers. If *A* cares about the disesteem of *B* and others like *B*, and they will disapprove someone who drinks the wrong beer, roots for the wrong team, goes to the wrong church, or dates the wrong people, then *A* may give up those things and conform to their preferences. He may do so even though conformity costs him more than they gain because he values avoiding their disesteem more than the consumption he forgoes. *A* is on balance better off if he avoids their disapproval, but in this case he would be still better off – and social welfare would be higher – if the others did not have the leverage of esteem.

As a final point, consider the link between esteem and observable traits (see McAdams 1995). Assume people care about and compete for esteem even from strangers. By definition, strangers have very little basis on which to make any esteem judgment. If *B* doesn't know anything about *A*, then *B* will presumably regard *A* as being just as worthy of esteem as an average human being. The only factors that may positively or negatively distinguish *A* in *B*'s eyes are matters that *B* can immediately observe while in *A*'s presence. Some of these will be *A*'s observed behavior. But another potential source of positive or negative distinction would be any observable traits, including those of sex and race.

Why would *B* esteem *A* differently merely because of *A*'s physical characteristics? The answer is that physical characteristics are used to define groups of individuals, and *B* may regard the average person of some groups as being more or less worthy of esteem than the average members of other groups. The point is obvious if we consider membership in organizations based on something other than physical characteristics. For example, *B* might have a strongly positive or negative evaluation of Yankee fans, military personnel, or members of a local golf club. Members of each group might wear clothing that reveals their group affiliation, thereby affecting *B*'s evaluation of them (perhaps intentionally). *B* now regards one of them with the esteem he has for average members of that group, even though if he knew more about the stranger he observes, he might regard them quite differently.

The same points apply when we consider groups of people sharing physical traits. *B* might believe that a particular observable characteristic correlates with other desirable or undesirable traits or behavior. *B* might reach this (correct or erroneous) belief based on his own observations and Bayesian reasoning, but the beliefs are more likely the result of relying on "conventional wisdom" that is the product of various cognitive and motivated biases. For example, *B* may believe that sex correlates with bravery or care-giving, which he esteems, and therefore regard the average woman and average man as deserving unequal levels of esteem. Given the way the biases operate, *B* is most likely to believe that, other things equal, his own observable physical traits correlate with positive character traits and behaviors and therefore

regard others who share such traits as being worthy of greater esteem than those who do not.

Once people believe such correlations exist, certain reinforcing processes come into play. I have discussed the relevance of physical traits from the perspective of a person giving esteem to a stranger. For the person *seeking* esteem from strangers, there are two kinds of strategies. One is to make sure that one's own public behavior is worthy of esteem. The other is to play the "group membership" game. That game involves many decisions. First, one can seek to join groups that enjoy high public regard and in which one's membership is publicly observable (i.e., because members use a visible marker not easily mimicked by non-members). Conversely, one wishes to exit groups that enjoy low regard and in which one's member is publicly observable. Second, in addition to entering and exiting groups, one can work to make one's high status memberships more visible and to make one's low status membership less visible.

Finally, holding constant one's memberships and their visibility, one can try to raise the average esteem accorded to the groups in which one's membership is visible. One way to do this is to engage in behavior, and to induce one's fellow members to engage in behavior, that will earn esteem from non-group members. But if, as I have claimed, esteem is (at least partly) relative, then another way to raise the esteem accorded one's group is to *lower* the esteem given to other groups. One may subordinate other groups by disparaging them and excluding their members from economically and socially productive exchanges. This strategy describes the practice of sex and race *discrimination* in, for example, employment, housing, or public accommodations. From the group's perspective, *A*'s costly effort to subordinate members of other groups may create a positive externality for other members of *A*'s group, who may then esteem *A* more for his contribution to group welfare. There is a lot to say about how these processes work (see McAdams 1995), but the result is a group norm of sex or race discrimination. Again, we see that arbitrary differences in observable personal traits can lead to a significant type of conformity.

In sum, coordination and esteem both produce conformity, and each is capable of causing conformity to behavioral patterns defined partly by the observable physical traits of individuals.*

Richard H. McAdams

University of Illinois

Notes

* I thank Dhammika Dharmapala, Chris Sanchirico, and Tom Ulen for comments on an earlier draft.

1. Moreover, it is unlikely the players would agree to a correlated equilibrium. In theory, player *A* might propose to player *B* to flip a coin and to assign to heads the outcome that *A* claims 100% and *B* claims 0%, and to assign to tails the outcome that *A* claims 0% and *B* claims 100%. The expected value for each player is then 50%, but if the players are at all risk averse, as is usually assumed, they will regard this highly variable outcome as worse than a certain share of 50% (which Skyrms' analysis shows they will likely achieve without randomizing).

2. More generally, the asymmetric strategies should include mixed strategies, where one claims *y* or *z* on average, with a variance around these means. For simplicity, I illustrate the point assuming the variance in each case is zero.

3. I.e., an equilibrium where females play Fair-minded against females and Greedy against males while males play Fair-minded against males and Modest against females. I will not continue to note these alternate possibilities, though they always exist. As I explain below, another theory is needed to explain which of these equilibria arise.

4. As a contrasting example, universal defection is an equilibrium in an iterated prisoners' dilemma even when cooperative equilibria are possible, but in an all-defect equilibrium, a given player prefers that others *not* play their equilibrium strategy. Therefore, an all-defect convention is not likely to become a norm.

REFERENCES

Aumann, Robert J. 1974 "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1, 67-96.

---------- 1987 "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1-18.

Brennan, Geoffrey & Philip Pettit 2003 *The Economy of Esteem*, Oxford: Oxford University Press.

Brown, Jennifer & Ian Ayres 1994 "Economic Rationales for Mediation," *Virginia Law Review*, 80, 323-395.

Coleman, James S. 1990 *Foundations of Social Theory*, Cambridge, MA: Harvard University Press.

Cowen, Tyler 2002 "The Esteem Theory of Norms," *Public Choice*, 113: 211-24.

D'Arms, Justin, Robert Batterman & Krzyzstof Gorny 1998 "Game Theoretic Explanations and the Evolution of Justice," *Philosophy of Science*, 65, 76-102.

Hadfield, Gillian K. 1999 "A Coordination Model of the Sexual Division of Labor," *Journal of Economic Behavior & Organization*, 40, 125-53.

Hirshleifer, Jack 1988 *Economic Behavior in Adversity*, Chicago: University of Chicago Press.

Lewis, David 1969 *Convention*, Cambridge, MA: Harvard University Press.

Mackie, Gerry 1996 "Ending Footbinding and Infibulation: A Convention Account," *American Sociological Review*, 61, 999-1017.

McAdams, Richard H. 1995 "Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination," *Harvard Law Review*, 108, 1003-1084.

---------- 1997 “The Origin, Development, and Regulation of Norms,” *Michigan Law Review*, 96, 338-443.

---------- 2000 “An Attitudinal Theory of Expressive Law,” *Oregon Law Review*, 79: 339-90.

---------- 2001 “Conventions and Norms: Philosophical Aspects,” in 4 *International Encyclopedia of the Social and Behavioral Sciences* 2735-41 (Elsevier Science, eds., N.J. Smelser & P.B. Baltes).

Nash, John 1950 “The Bargaining Problem,” *Econometrica* 18, 155-62.

Pettit, Philip 1990 “*Virtus Normativa*: Rational Choice Perspectives,” *Ethics*, 100, 725-55.

Skyrms, Brian 1994 “Sex and Justice,” *The Journal of Philosophy*, 91, 305-20.

---------- 1996 *Evolution of the Social Contract*, New York: Cambridge University Press.

Sugden, Robert 1986 *The Economics of Rights, Co-operation and Welfare*, Oxford: Blackwell.

---------- 1990 “Contractarianism and Norms,” *Ethics*, 100, 768-786.

---------- 1998 “Normative Expectations: The Simultaneous Evolution of Institutions and Norms,” in *Economics, Values, and Organization* 73 (Avner Ben-Ner & Louis Putterman eds., Cambridge: Cambridge University Press).

Ullmann-Margalit, Edna 1977 *The Emergence of Norms*, Oxford: Oxford University Press.

Young, H. Peyton 1998 *Individual Strategy and Social Structure*, Princeton: Princeton University Press.