EXPRESSIVE LAW AND OPPRESSIVE NORMS: A COMMENT ON RICHARD McADAMS'S "A FOCAL POINT THEORY OF EXPRESSIVE LAW"

Amy L. Wax[*]

I. THE PROBLEM OF "OPPRESSIVE" NORMS

HOW does law affect behavior? Scholars are not of one mind. The conventional "deterrence" or "instrumental" model[1] looks to the response to penalties and punishments imposed for violations of legal rules. People's willingness to comply with the law's commands stems from official threats to deprive them of something valuable by imposing financial liability, withdrawing privileges, throwing them in prison, or extracting fines.[2] The simplicity and testability of this theory account for its appeal and for the temptation to adopt its assumptions in analyzing particular problems.

Another line of scholarship recognizes the model's limitations. A theory of compliance that looks only to the response to sanctions and penalties cannot explain observed patterns of behavior. The scattershot nature of law enforcement predicts more widespread flouting of legal commands than is actually observed. It is fortunate, however, that the deterrence theory fails to explain why most people obey the law. Were a credible threat of punishment all that worked to prevent violations of law, vastly greater investments in law enforcement and penalties would be needed to secure a tolerably orderly society.

The recent surge of interest in the expressive power of law represents an attempt to develop a more complete and sophisti-

[*] Class of 1948 Professor of Scholarly Research in Law, University of Virginia Law School. Thanks to Chris Sanchirico and Paul Mahoney for excellent comments and suggestions.

[1] See Tom R. Tyler, Why People Obey the Law (1990).

[2] See, e.g., id. at 3 (arguing that the instrumental perspective views people "as shaping their behavior to respond to changes in the tangible, immediate incentives and penalties associated with following the law—[with] judgments about the personal gains and losses resulting from different kinds of behavior").

cated view of how law affects behavior. That the law can induce compliance wholly apart from its power to punish has strong intuitive appeal and comports with the evidence that a simple incentive model cannot explain everything. The challenge, however, is to develop a plausible theory of how legal rules influence behavior. This requires a more sophisticated picture of human motivation and interaction than is ordinarily presented by the instrumental view of law.

Richard McAdams's Article, "A Focal Point Theory of Expressive Law,"[3] represents a useful contribution to this effort. Key to his analysis is the understanding that the penalties the state imposes are not the only factors that influence individual reactions to legal rules. What government can do to us is often less important than what others will do to us in response to official pronouncements. Therefore, in calculating the costs and benefits of a particular action, people must take into account other persons' reactions to laws as well. What game theory adds to the instrumental model, then, is not a fundamental change in the assumption of self-interested behavior, but a more sophisticated framework for predicting law's effects based on a citizen's complex interactions with others.

Because of its usefulness in modeling a variety of social interactions that the law might seek to regulate, the so-called "Chicken," or Hawk-Dove game dominates McAdams's discussion.[4] This game captures the dynamics of many situations in which parties can benefit from cooperation but have incompatible claims to valuable resources. The hallmark of Hawk-Dove interactions is a payoff

[3] 86 Va. L. Rev. 1649 (2000).

[4] It looms large for other legal commentators as well. See, e.g., Eric Rasmusen, Games and Information: An Introduction to Game Theory (1989); Robert Sugden, The Economics of Rights, Co-operation and Welfare 58–62 (1986); Joan Williams, *Un*bending Gender (2000); Carol M. Rose, Women and Property: Gaining and Losing, 78 Va. L. Rev. 421, 428 (1992) (referring to a zero sum game); Amy L. Wax, Bargaining in the Shadow of the Market: Is there a Future for Egalitarian Marriage?, 84 Va. L. Rev. 509, 556–57 (1998); Paul Mahoney & Chris Sanchirico, Competing Norms and Social Evolution: Is the Fittest Norm Efficient? (May 1, 2000) (unpublished manuscript, on file with the Virginia Law Review Association), available on the Social Science Research Network, http://www.ssrn.com, as Legal Studies Working Paper No. 00-15; Randal C. Picker, Endogenous Neighborhoods and Norms (Feb. 7, 2000) (unpublished manuscript, on file with the Virginia Law Review Association); Eric Posner, The Evolution of Constitutions 5 (Feb. 28, 2000) (unpublished manuscript, on file with the Virginia Law Review Association).

scheme that generates a characteristic order of preference for combinations of moves. The first choice for each player is to assert his claim (dominate or play Hawk) while the other yields (submits or plays Dove). Barring that, a player prefers to yield while the other does as well (that is both play Dove). Next in line is deferring to the other player (the Dove/Hawk combination). Dead last is the prospect of conflict, in which both assert their claim simultaneously. Because conflict results in the smallest gain or the greatest loss for each player compared to other combinations, it is best avoided.

The expressive role for law put forward by McAdams depends crucially on a key characteristic of the types of Hawk-Dove games that are the centerpiece of his Article. In these games, there is no single "pure" strategy of play that is best for all players and from which no player has an incentive to deviate. There is no unique Nash equilibrium because there is no one uniform strategy for each player that is the best reply to itself.[5] By creating a focal point, or "asymmetry" as McAdams calls it, around which players can coordinate their moves, the law helps private actors settle into one "pure-strategy" equilibrium for playing the game. That equilibrium enables players to avoid falling into the destructive "fourth box," or Hawk/Hawk combination, which is costly for everyone. To the extent that law can foster the emergence of a stable nonconflictual equilibrium, it helps increase net social welfare.

McAdams's theory of expressive law depends on the law's ability to foster expectations in players' minds about how others will move. Because each player stands to gain if he coordinates his actions with others, the ability to predict how opponents will behave confers an advantage. According to McAdams, the law imparts useful information through its capacity to call attention to specific options for playing the game. As an empirical matter, highlighting

[5] See Rasmusen, supra note 4, at 73 (explaining that the simple chicken game has no unique Nash solution; rather it has a mixed strategy equilibrium and "two pure-strategy Nash equilibria"). This statement is an oversimplification. Strictly speaking, there may be one Nash solution in a few cases where payoffs differ for row and column players, because there may be some games in which there is a strictly dominant (although different) strategy for each player. Nonetheless, the existence of more than one equilibrium strategy will generally be assumed for games under consideration in this Comment, including the games with different row and column payoffs discussed herein.

an option creates the expectation that everyone will follow that option. Because each player strives to coordinate with others, each willingly takes a cue from the expectations the law creates. Thus does law as expression harness individual self-interest to influence action in desirable directions without imposing sanctions on anyone.

The purpose of this Comment is not to question the basic elements of McAdams's account of how law affects expectations and players' choice of strategy. Rather, its goal is to apply and extend his insights to a variation on the basic game theoretic model that is the focus of his analysis. The examples McAdams uses to develop his theory represent the simplest types of Hawk-Dove situations. These games share two key features that are crucial to the discussion in this Comment. First, as exemplified by Figure 4 in McAdams's Article,[6] they display a characteristic pattern of payoffs in which each player receives the same amount for a particular response to an opponent's move regardless of whether that player occupies the row (designated by McAdams as *R*) or column (designated by McAdams as *C*) position in the game. For purposes of this Comment, arrays fitting this pattern are designated "balanced arrays." Games that deviate from this pattern by assigning different payoffs to *R* and *C* players for the same strategic combinations —as in the examples provided in the Appendix to this Comment—will be designated "unbalanced" games.

The second feature shared by most games in McAdams's Article is a high degree of mobility or interchangeability between roles in the game. In reiterated games with repetitive rounds of play, a player occupying the row position on one round is eligible to occupy the column position on another. Games in which players can easily swap roles will be designated "fluid" or "nonrigid" and those in which they cannot will be designated "fixed" or "rigid."

The purpose of this Comment is to explore the implications of McAdams's theory of expressive law for Hawk-Dove games in which payoffs are unbalanced and role assignments fairly fixed. These games are of special interest because many social interactions can be modeled along these lines and those interactions are often at issue in troubling conflicts that attend various forms of so-

[6] McAdams, supra note 3, at 1675.

cial inequality. In particular, the conventions observed in these circumstances tend to be "oppressive" in giving rise to durable inequalities among identifiable social groups. Taking for granted the basic features of McAdams's analysis, this Comment will examine whether and how law in its expressive capacity might operate to influence the "oppressive" conventions that tend to emerge in social situations that mimic unbalanced games with rigid roles and how the law might alter those conventions once established.

The potential to generate oppressive conventions is not unique to games with differential payoffs, however: Oppressive norms can emerge in balanced as well as unbalanced games. Within the two pure-strategy equilibrium options (Hawk/Dove or Dove/Hawk), players assigned the aggressive role (Hawk) will do better than those assigned the passive role (Dove) even when primary payoffs are the same for everyone.[7] This shows that the "oppressiveness" of norms is primarily a function of the fixity or fluidity of roles rather than of the values in the payoff array. The emergence of stable conventions will consign those stuck playing the nonaggressive position to permanent disadvantage unless they are able to "change places" with opponents with ease. The inequalities of payoffs that characterize "pure-strategy" equilibria in Hawk-Dove games thus threaten to create a permanent "overclass" and "underclass" of individuals—and to become potentially problematic—only when there is little mobility among players occupying the different positions in the game.

This Comment argues that there are features of "oppressive" conventions that tend to emerge in games with lopsided payoffs that render them of special interest and concern. Those features are also the very ones that pose special challenges to McAdams's theory of expressive law because they predict resistance to influence or centralized orchestration through the powers McAdams assigns to legal expression. First, as explained more fully below, many of the games that fit this pattern and are of the greatest social interest revolve around "naturally" salient characteristics of persons, such as race, sex, or ethnicity. Play will tend spontaneously to coordinate around these features, obviating any role for a legally suggested or mandated focal point. Second, the conventions that

[7] See id. at 1674–75.

dominate these types of games may be particularly resistant to official influence at the formation stage because they are driven by individual responses to payoff values. Third, oppressive norms in unbalanced games would appear to be particularly impervious to alteration by legal expression, because the risks from attempting to modify the dominant conventions are greatest for players who stand to benefit most, and law, apart from its ability to sanction, would not appear to alter those risks.

Yet such norms do change. Although McAdams's discussion is concerned both with the role of expressive law in fostering the emergence of "pure" strategy conventions and in how law can change conventions once they emerge, the discussion here will center on how law in this capacity might operate to fuel shifts in "oppressive" conventions that tend to emerge in unbalanced games. The reasons for concentrating on norm change in this context are twofold. First, because the informal forces driving the games at issue towards particular equilibria are strong, the problem in many Hawk-Dove-like social circumstances is not the absence of coordinated play or imperfect coordination but a high degree of coordinated play in socially problematic directions. Second, permanent and sweeping social inequalities, productive of widespread disaffection and social unrest, may result from the entrenchment of "oppressive" norms. These consequences will tend to spur agitation to overturn existing conventions or mitigate their outcomes. An understanding of a potential role for expressive law in fostering change is vital to the project of altering these norms.

In explaining how norm changes might occur in these contexts, this Comment rejects McAdams's account of norm change without denying the possibility that law plays an expressive role. Instead it offers a different theory of how legal expression might contribute to the defeat of "oppressive" norms. Specifically, the account offered here minimizes the importance of "cranks" (individuals who suffer minimally from conflict), to whom McAdams assigns a pivotal role in norm change. It relies instead on morally outraged underdogs (or those assigned the nonaggressive role under the status quo). It predicts that underdogs' self-sacrificial moves, as spurred by moral sentiments and encouraged by law, will fuel strategic shifts towards conventions that favor those disadvantaged by previously entrenched norms.

II. VARIATIONS ON A THEME: UNBALANCED PAYOFFS AND FIXED ROLES

A. Unbalanced Payoffs

It is best to begin by considering in more detail the peculiar features of the games that dominate McAdams's Article. As suggested, the payoff array in Figure 4 is typical. The first feature of importance is that the expected payoff from playing Hawk or Dove, holding the other player's move constant, is the same for each player regardless of whether that player occupies the row (*R*) or column (C) position. If any player plays Dove to another's Dove, he gains a payoff of 1. If he plays Dove to the other's Hawk, he gains nothing. Playing Hawk to another Hawk always produces a loss of 2, and Hawk to Dove always earns 2 regardless of who plays that role. Thus, the gains from adopting an aggressive or submissive posture—from being assertive or deferential—are the same for everyone, and do not depend on identity, social role, or assigned position within the game. This condition produces the classic balanced array for the Hawk-Dove game. That array is symmetrical within the matched strategy boxes (the gains to playing Hawk match those to Hawk for the Hawk/Hawk combination and the gains to Dove are the same for both players in Dove/Dove). It is also symmetrical across the diagonal for mismatched strategies (the Hawk/Dove payoffs are the flip side of the Dove/Hawk payoffs). Because McAdams uses the terms "symmetrical" and "asymmetrical" to refer to pure-strategy conventions that emerge with "labeled" or focal-point play, rather than to the pattern of primary payoffs within the game, this Comment will refrain from using that terminology to describe payoff patterns. Rather, it will stick to the term "balanced" or "unbalanced" to describe arrays similar, respectively, to that in Figure 4 of McAdams's Article, or to arrays similar to those in the games set forth in the Appendix to this Comment.

In contrast, an unbalanced Hawk-Dove game is one in which row and column players can expect to fare very differently from opponents depending on whether they adopt an aggressive or nonaggressive posture. The following (see Game #1 in Appendix, p. 1777) provides an example of such an array:

		B	
		Dove	*Hawk*
R	*Dove*	8, 5	5, 15
	Hawk	-12, 2	-10,-2

There are many possible variations on this basic theme with complex implications for strategies of play. For example, unbalanced games can arguably be divided into two types. There are those games in which the actions that constitute the aggressive and passive choices are precisely the same for both players (for example, fighting for a piece of property versus backing off), but payoffs differ for different players due to individual attributes or circumstances. The Division of Labor game (see Appendix, p. 1778) is one such game. In other games, the game is formally structured so that each party's options can be viewed as either Hawklike or Dovelike (that is, assertive or acquiescent), but the parties do not actually make identical moves within the game. The sexual harassment game is one such example. Since the strategy options within the game for men and women are not precisely the same,[8] there is no reason to think the payoffs would be either. Despite the differences among these games, the Hawk-Dove model can serve for interactions with differential payoffs in which the Hawklike or Dovelike actions for players are either the same or not.

What these games have in common with each other and with the classic balanced Hawk-Dove games is the parties' order of preference. Each player prefers the dominant position (playing Hawk to the other's Dove) to the submissive one. Cooperation (Dove/Dove) is second best and conflict is the least desirable for each. The games set forth here generally differ from the classic balanced array in that the column player (*C*) has more to gain overall from being assertive and playing Hawk than does his opponent, regardless of what the other party does. Likewise, player *R*'s expected gains from playing Dove are greater than *C*'s. In addition, although both parties have a paramount interest in avoiding conflict, *C* has less to

[8] We can say that the man is playing the "Hawk" strategy by engaging in harassment. But the Hawkish woman doesn't harass back—she does something different, like trying to get her harasser fired. So even though both moves are "Hawklike"—in that they deliver the largest possible payoff to each player when the other player backs down—they are clearly very different actions.

lose than his opponent from the conflictual situation (Hawk/Hawk). Variations on this pattern are possible within the context of a range of payoffs for row and column players. For example, *C* might have more to gain from playing Hawk to Dove than *R*, but more to lose (or less to gain) from conflict. These variations may give rise to different predictions about which equilibria will emerge spontaneously or in response to outside cues. For purposes of this Comment, discussion will be confined to the general pattern exemplified by the subset of games set forth in the Appendix, which are arguably typical of at least some real-life situations.[9]

B. Fixed Roles

The other pertinent feature of the games that dominate McAdams's discussion is that players' roles are more or less interchangeable. Each player is free on successive rounds to take turns occupying the row or column position at will. Alternatively, circumstances guarantee that a player will find himself in each position on a fairly random basis. As already noted, interchange-

[9] How might games with unbalanced payoffs emerge in social situations? What are the real-life psychological analogs of the costs and benefits represented by these characteristic arrays? The payoffs in the sample game above, for example, would result if Ms. *R* had a greater taste for cooperation than Mr. *C* so that, regardless of her opponent's move, playing Dove would always be less painful or more pleasant for her than for him. Alternatively, *R* may assign independent importance to harmonious and cooperative relations wholly apart from any tangible benefits it might bring, whereas *C* might care less about social harmony or vaguely disdain it. For this reason, *R* might gain more psychic utility from playing Dove to an opponents' Dove. Along the same lines, *C* might relish dominance or dislike playing the submissive role, thus gaining extra psychic utility from playing Hawk to Dove and less benefit from adopting the deferential posture. *R* might have the opposite preferences: Although she would enjoy the tangible benefits of playing the dominant role, those gains must be discounted by her greater discomfort with the Hawk role and lesser distaste for submission. *R* might also harbor a more intense aversion to conflict as such relative to *C*, so that she suffers greater costs from finding herself caught in a Hawk/Hawk situation and is more eager to avoid it. Alternatively, *R* may have more to lose from conflict than her opponent because he is stronger, bigger, or more wealthy and can "hurt her more" if they come to blows. *R* may have fewer and less desirable alternatives than *C* outside the interactive context, which would be reflected in the overall pattern of payoffs in the array.

As this discussion suggests, these types of intraplayer differences may characterize the interactions of women and men at home and at work. Social relations among members of different cultural, ethnic, or religious groups could also be modeled this way. For discussion in the context of gender, see, for example, Nancy Folbre & Thomas E. Weisskopf, Did Father Know Best? Families, Markets and the Supply of Caring Labor, *in* Economics, Values and Organizations 171, 188 (Avner Ben-Ner & Louis Putterman eds., 1998); Rose, supra note 4; Wax, supra note 4.

ability matters whether arrays are balanced or unbalanced. If the focal feature that correlates with the assertive or yielding role within an established pure-strategy convention is one that picks out a fixed subgroup, one category of players can enjoy enduring gains at the expense of the other. That can happen when the chosen feature is immutable or when it is linked to circumstances or traits that are difficult or costly to modify. The convention that blue-eyed drivers yield to brown-eyed drivers may create a world in which the blue-eyed are chronically late and the brown-eyed always on time. That inequality does not depend on different payoffs assigned to different players in the array, but only on the emergence of an equilibrium that ties moves in the game to an inborn trait.

McAdams's discussion is primarily concerned with games where the prospect of winning or losing under the dominant convention is not fixed forever. In the first possession or "property" game as described by McAdams,[10] it is open to most individuals within society to acquire property and defend it, taking a turn at playing Hawk to someone else's Dove. Yet property owners will also find themselves respecting others' property rights, thereby playing Doves to others' Hawks. In the crossroads game, the traveler approaching the intersection from the "wrong" direction (left or right, or green pole side or opposite) must defer to someone approaching at right angles and bear with the inconvenience and aggravation of delay. But chances are that the next time that same traveler will approach from the opposite direction and will be allowed to go first. The smoker/nonsmoker game is less clear-cut, but experience suggests that the roles there are fairly mutable as well. A smoker today has the chance to become a nonsmoker tomorrow and vice versa. Although smokers may find it hard to quit and some nonsmokers may never be tempted, it is not uncommon for people to resume smoking or quit smoking more than once in a lifetime.

Fluidity of roles is not the exclusive preserve of balanced games. Indeed, balanced payoffs are the exception rather than the rule for joint action in real life. Most interactions—whether between fixed or fluid categories of individuals—do not generate identical gains for all participants. Employers and employees, fellow employees,

[10] McAdams, supra note 3, at 1693–95; see also Sugden, supra note 4, at 70–71 (describing how the Hawk-Dove game is likely to lead to some convention of property).

buyers and sellers, business partners, roommates, marital partners, sharers of public goods, and users of public accommodations all cooperate in ventures where unequal payoffs are the rule rather than the exception. The fate of participants is not absolutely fixed, however, because there is usually some potential for moving from one role into another. As noted, most people have the wherewithal to own property. An employee who works for others may realize his dream of starting his own business and getting others to work for him. Many enterprises function simultaneously as buyers and sellers, and people who are galled by the rules for buyers can move into sales. In numerous contexts, parties unhappy with conventions that disfavor their kind can take steps to transform themselves into the opposition.

Nonetheless, swapping roles within games of conflict and cooperation is rarely without costs. Those costs are not uniform across contexts and can vary considerably among individual social actors. Fluidity and interchangeability lie on a continuum, and the burdens of switching positions are often not the same in both directions for any game. It is harder to become a nonsmoker than a smoker, easier to lose one's property than to obtain and keep it. At one end of the continuum lie true "equal opportunity" games—those in which players enjoy virtually unimpeded access to alternative roles and freely assume them. At the other end are those that assign roles based either on immutable traits or on roles so costly to change—because of large sunk costs or the need for fresh heavy investments—that they are effectively immutable in practice.

III. UNBALANCED PAYOFFS AND FIXED ROLES: THE WORKINGS OF EXPRESSIVE LAW

Although the potential for oppressiveness is at least theoretically present in both balanced and unbalanced games, the remainder of this Comment will be concerned primarily with games characterized by *both* an imbalance in payoffs *and* fixity of roles. The discussion will investigate the potential role of legal expression in the formation and alteration of conventions that arise in these games. The quest to understand the place of law as expression in this context is of theoretical as well as practical concern. The examples set forth in the Appendix attempt to model well-known interactions between persons of different races and sexes. These

games tend to gravitate towards conventions that have important distributional and social consequences for the fate of these social groups. The dynamics of the Hawk-Dove game predict that members of one group will repeatedly tend to play Hawk to the other group's Dove, thereby garnering the lion's share of the gains from cooperation. Members of the group who play Doves to the others' Hawks will find themselves at a permanent disadvantage. Because there is little potential for "trading places," the consequences of playing the game will not even out. These are the settings in which the emergence of "oppressive" norms appears most troublesome and perplexing and poses the greatest challenge to law's capacity to change private patterns of behavior.

How might the law influence the conventions that emerge for the conduct of games typified by the Appendix examples? How might the law change those conventions once they become entrenched? This discussion suggests that there are two key elements of these games that have important implications for the process of norm formation and for law's influence on that process. The first is the "natural" salience of the particular focal features around which strategies in these games tend to coordinate in practice; the second is the manner in which the values in the unbalanced payoff arrays tend to influence the conventions that emerge. These issues will be examined in the remainder of Part III. The fundamental structure of these games also affects the options for changing those norms that have already become established. In Parts IV and V of this Comment, I will set forth a theory of norm change and of expressive law's place within it that represents an alternative to McAdams's thesis. The proposed account fits better with observed social patterns and is tailored more precisely to the specific dynamics of the games under consideration here.

A. "Natural" Salience and Labeling Features for Coordination

McAdams identifies one initial function for law in norm formation as the selection and advertisement of an "asymmetry" or focal point around which pure-strategy equilibria can converge.[11] The options are to create an artificial feature, such as green poles on unmarked pavement, or to draw attention to some existing element

[11] McAdams, supra note 3, at Section I.C.1.

of the world or of social life. The characteristic chosen may have some degree of "natural" salience—it may be something that people tend to notice spontaneously, Or it may be some arbitrary element of the environment that is not particularly interesting in itself but that comes to command attention just because the law picks it out for special notice. McAdams wonders more than once why and how the observable features (or "natural asymmetries") around which pure-strategy conventions actually coordinate are selected from the myriad ones available around us.[12] Clearly, many such features come into use without official intervention, and the corresponding conventions appear to arise spontaneously without significant prodding from centralized authorities.[13] Alternatively, legal rules may "piggyback" on preexisting tendencies, and amplify spontaneous processes, by picking out naturally salient features as focal points for coordination. Although McAdams acknowledges that many conventions arise informally though extralegal mechanisms, he nevertheless suggests that there remains a significant role for law in fostering coordination and encouraging the emergence of pure equilibria in many cases. Critical to that possibility is the law's deliberate selection and advertisement of focal features or asymmetries around which pure strategies can coalesce.

The emergence of oppressive pure-strategy norms in the games of interest here depends critically on players' tending to notice certain elements like race and sex and to make them the basis for coordinated strategies of play. The ubiquity of race- and sex-based conventions in the conduct of social life worldwide and throughout history suggests that these features are "naturally" salient. As clear, public, universal, and easily discernible parameters, sex and race may have special psychological appeal or may be especially well suited for coordinating social action. In addition, the practical need to identify, respond to, cooperate with, or defend against individuals from disparate groups or the opposite sex provides a compelling basis for a "natural" interest in racial and sexual characteristics. For all these reasons, racial and sexual differences are likely to emerge spontaneously as focal points in the types of coor-

[12] See id. at 1693–96.

[13] See Posner, supra note 4, at 30 (discussing the idea that common law follows on, rather than creates, conventions of property and social interaction that evolved informally).

dination and conflict games that can be played more efficiently in pure-strategy equilibria.

McAdams's analytic framework helps make sense of the pervasiveness of sex- and race-based conventions and roles. If sex and race are naturally more salient and more interesting than green poles and many other characteristics, pure-strategy equilibria tied to those attributes can be expected to emerge extralegally, spontaneously, and informally on a fairly regular basis. But to the extent that "labeling" is the vital first step—and may be the only important step—in initiating the dynamic that leads to pure coordinated play, the self-executing choice of race or sex as a focal point of coordination suggests a minimal role for law in norm formation. Since, by hypothesis, labeling is most of what the law does,[14] this leaves the law with little work to do in fostering coordinated strategies for familiar interactions involving men and women or members of different groups. That sex- and group-identity-based conventions tend to emerge repeatedly in different cultures worldwide suggests that official attempts to select and advertise focal points have not played a major role in the emergence of many important practices that dominate informal social interactions. Official nudging in these contexts may be both unnecessary and irrelevant.[15]

B. Convention Formation

Although McAdams suggests that "labeling" or picking out a focal feature for coordination is the critical beneficial function for expressive law in fostering salutary social cooperation,[16] that is not all the law does in many cases. Selecting a feature for coordinated play does not in itself determine which of two pure-strategy equilibrium options will emerge as the dominant convention. The choice potentially depends on a number of factors that operate in different contexts. Chance, random variation, "bounded rationality," mistakes, or experimentation[17] may push play towards one

[14] But see infra Section III.B.

[15] McAdams's analysis (and this Comment's gloss on it) suggests why the belief that law has been central to "constructing" sex- or race-based social rules—if not in helping to sustain them—may be unwarranted.

[16] McAdams, supra note 3, 1691–95.

[17] See sources cited infra note 26.

pure strategy or the other. Or a preexisting expectation—perhaps from some intrinsic psychological tendency—may link one value of the coordinating feature to an aggressive or submissive role. For example, where the location of property operates as a salient feature in the "possession" game, all players might share the expectation that the player closest to the property will aggressively claim it. Or people of both sexes may expect that the person physically closest to the baby will care for it. Although the reasons are not always well understood, the direction of players' responses to a salient feature will not always be random.

Alternatively, the law can try to choose the prevailing convention by announcing a particular strategy for play, such as when it declares that those on the side of the green poles must yield. The argument here, however, is that the law will have less leeway to choose one rule rather than another where payoffs are not the same for all players. Players in unbalanced games will tend to move inexorably and spontaneously toward one pure-strategy equilibrium rather than the other because the structure of the payoffs is biased toward the emergence of one conventional mode of play.

1. Emergence of Conventions in Balanced Games

Legal authorities in charge of traffic can set up green poles by roadsides and call attention to them. Although coordination may result with time, the actual rule that will emerge cannot be predicted. It is as likely that one green-pole rule (pole-side yield) will emerge as it is that its opposite (pole-side right-of-way) will. Alternatively, the law can go further and declare that vehicles approaching an intersection on the green pole side shall have the right of way. McAdams suggests that, even absent direct enforcement, the mere declaration makes it more likely that the announced rule will prevail.[18]

In most games with a balanced payoff structure, the central authority will have no reason to favor one rule over another: It matters only that there be a rule. The principal purpose is to eliminate costly conflicts, and this can be accomplished regardless of which convention wins out. One convention will not ordinarily be more socially efficient or beneficial overall, since either produces

[18] See McAdams, supra note 3, at 1681–83.

the same *sum total* of payoffs to the players. Although one convention might result in the "oppression" of one group by another (as when coordination is tied to some fixed identity trait), switching conventions is, in this context, a zero-sum game.[19] Alternatively, the existence of negative externalities or undesirable third-party effects may provide a reason to favor one pure-strategy convention over its opposite.[20] Absent such conditions, however, the equality of total payoffs from the game regardless of which convention is chosen will make the choice between them a matter of indifference, if not to the players themselves, then to the social unit overall.

As for the players in balanced games, they definitely will favor the convention that favors them. But absent some preexisting bias, psychological or otherwise, shared by all,[21] individual players cannot do much, consistent with rational self-interest, to push conventions in a personally beneficial direction. As McAdams's analysis makes clear, the payoff structure of balanced games means that the convention that emerges from "labeled" play in the Hawk-Dove context depends largely on chance perturbations and arbitrary deviations from randomness. Self-seeking responses to those variations or mistakes cause conventions to evolve over time in one direction or another.[22] But the similarity of payoffs to all players means that the direction of evolution cannot be predicted ahead of time. There is no bias *inherent* in the structure of the game that favors evolving towards one convention over the other.

Sex roles are an interesting case in point. Even if payoffs were similar for men and women in their cooperative interactions—and

[19] See discussion infra pp. 1750–51 on "table-turning" and changing oppressive norms.

[20] An example is the smoker/nonsmoker game. Although there may be no reasons intrinsic to the game to choose the nonsmoker-dominant over the smoker-dominant norm, there may be external reasons to do so: the costs to third parties, not reflected in direct losses to smokers, of increased death and morbidity from smoking. See discussion infra Section IV.B. Other balanced games giving rise to norms that track immutable traits, by creating entrenched advantages and disadvantages, might also generate undesirable extrinsic social consequences. But it is hard to see why anyone outside the game would care whether those driving on the green-pole side have the right of way or not.

[21] See discussion supra Section III.A.

[22] See, e.g., Mahoney & Sanchirico, supra note 4 (describing evolution by mutation); Picker, supra note 4 (describing how movements by different agents change eventual outcomes).

were indeed the same for all potential players regardless of sex—conventions might still emerge in which dominant and submissive roles correlated with sex. The natural salience of sex means that the assignment of social roles by sex—as in the "division of labor game" in the Appendix—could still occur even in the absence of any sex difference in preferences or payoffs. But if the array in sexual "division of labor" games in different cultures were indeed balanced in every case, that should give rise to very different social and historical patterns than those observed. Because there would then be nothing in the game that favored one equilibrium convention over its opposite, men should end up playing the dominant role under the conventions that emerge in some times and places and women should dominate in others.[23] One would not expect to see similar tasks assigned to each sex in most societies in which the game was played.

2. Emergence of Conventions in Unbalanced Games

This is not the pattern of sex role differentiation that is actually observed. Historical and cultural reality are more consistent with unbalanced payoffs in the division of labor game. Although "labeling" creates the potential to move toward one of two pure-strategy equilibrium conventions in games with unbalanced arrays, these conventions are not equally likely to evolve under the pressure of the players' spontaneous choices. Unlike with the balanced array, the expected payoff from any move (Hawk or Dove) is not the same for each player.[24] In Game #1 (Appendix, p. 1777) for example, player *R* has less reason to take assertive action than *C*. Assuming an initially equal chance that an opponent will fight or back down, *R* gains less than *C* from dominating his opponent and suffers a greater loss from conflict. Not only does *R* have less to gain from being assertive than his opponent, but assertiveness (playing Hawk) has a lower expected value for him than submis-

[23] Once again, this analysis makes the simplifying assumption that there are no other preexisting factors that tend to tilt the convention in one direction or that establish a shared psychological link between certain players and particular roles or tasks. See, e.g., McAdams, supra note 3, at 1663 (speculating that rationality cannot "fully explain why some solutions 'stick out' from others").

[24] *C* has an expected value from Hawk of 0.5 * (15 – 2) = 6.5. *R* has an expected value from Hawk of 0.5 * (12 – 10) = 1. *C* has an expected value from Dove of 0.5 * (5 + 2) = 3.5. *R* has an expected value from Dove of 0.5 * (8 + 5) = 6.5.

sion (playing Dove). Since *R* gains less from dominance, he will tend to "try out" the nonaggressive posture (Dove) more often on multiple rounds of play. Likewise, because *C* expects to gain more from dominance than from playing Dove, he will opt for an aggressive posture more often. This reflects the high gains to *C* from dominating *R* as well as the smaller downside risk to *C* from conflict. Thus even knowing nothing of their opponents' payoffs, expectations, and preferences, *R* and *C* will make a very different mix of moves in their initial forays into the game.[25] *C* will tend to favor the Hawk posture, and *R* will tend to choose Dove. But the more often *C* plays Hawk, the more often it behooves *R* to respond by playing Dove (to avoid conflict), and vice versa. Since Hawk is *C*'s best response to *R*'s Dove, *C* will play Hawk with increasing frequency. Play will move swiftly towards the equilibrium in which *R* plays the "underdog" and *C* the "top dog." In sum, the individual choices players can be expected to make at the outset will tend to favor the emergence of only one of the two available Nash equilibria as the dominant convention for play. That bias results from players' responses to their own expected payoffs, which are implicit in the structure of the game. Moreover, if players' roles have payoff-driven tendencies, then conventions in which they go with those tendencies ("natural" conventions) will tend to be difficult to unseat (via "mutation") relative to conventions in which players contravene their tendencies.[26] This means that most of the time the natural convention will be in effect.

[25] Players' knowing something of their opponents' payoffs will exacerbate the tendency to make disparate moves at the outset and help speed progress towards the *C*-dominant/*R*-submissive equilibrium. Thus, if *R* knows that *C* has a lot to gain by playing Hawk, *R* will, at the margin, believe that *C* is more likely to play Hawk. It will then pay for *R* to play Dove most of the time, which will establish the reigning equilibrium even more quickly than if *R* had no preexisting knowledge of *C*'s payoffs but knew only his own.

[26] Another way to see this point is through the concept of the "tippability" of equilibrium conventions, as explored most recently by Paul Mahoney and Chris Sanchirico. See Mahoney & Sanchirico, supra note 4, at 21–24. For more on this, see Jonathan Bendor & Piotr Swistak, The Evolutionary Stability of Cooperation, 91 Am. Pol. Sci. Rev. 290 (1997); Dean Foster & Peyton Young, Stochastic Evolutionary Game Dynamics, 38 Theoretical Population Biology 219 (1990); Michihiro Kandori et al., Learning, Mutation, and Long Run Equilibria in Games, 61 Econometrica 29 (1993); H. Peyton Young, The Evolution of Conventions, 61 Econometrica 57 (1993). Assuming that players sometimes deviate randomly by "mutation" or otherwise from the moves dictated by the entrenched equilibrium convention, the question is how many mutations or deviations are required to make it in everyone's self-interest to move to the alternative stable convention. In our case, for example, assume that the

What are the implications of these observations for the role of expressive law in regulating conventions in unbalanced games? McAdams devotes a good deal of attention to law's power to spur the emergence of efficient pure-strategy conventions out of the less efficient chaos of mixed strategy or nonconventional play. Law can do good at low cost by creating focality. In addition, it can speed progress toward coordinated play by selecting a particular convention or suggesting a particular rule. These are elegant and important insights. But the account of the evolution of conventions in unbalanced games set out here points up the limitations inherent in this picture of law's expressive role. In this setting the choice of convention is not up for grabs. Rather, players spontaneously make moves that drive the game forcefully towards one convention rather than the other. If the structure of payoffs means that *R* generally chooses to play Dove more often than *C* plays Hawk, it is in both players' interest to continue on the path towards the pure equilibrium in which *C* is the aggressor and *R* submits. Not only does that progression need little prodding from law, but it is hard to imagine how law in its expressive capacity (that is, apart from altering the payoffs in the array itself) could operate to influence play away from the predicted direction. Unlike the balanced case, where disarray can persist for a long time because "nothing favors a particular pure Nash equilibrium,"[27] disarray is unlikely to persist for very long where payoffs diverge. The interests of the players take over to drive away disarray and enshrine the expected conventions. This account suggests that, as a practical matter, there is very little room for law as expression to operate in shaping the ini-

existing convention is the opposite of what we predict would emerge: that is, assume that all column (*C*) players play Dove and all row (*R*) players play Hawk. Tipping this "unnatural" convention would require relatively few mutations. That is because it would not take many encounters with a "mutant" *R* player (playing "natural" Dove instead of his "unnatural" Hawk role) to make it in the interest of *C* players to convert to Hawk, and vice versa. Players do so much better with the *C*-Hawk/*R*-Dove convention than with the opposite that it does not take many forays into the former combination for it to emerge decisively as the favored one.

In contrast, the "natural" convention of *C*-Hawk/*R*-Dove is difficult to unseat through the process of mutation. Intrinsic payoffs generally make *R* players reluctant to play Hawk and *C* players reluctant to play Dove. Starting from the "natural" convention, therefore, there would have to be a lot of "mutant" *R* players switching to Hawk before *C* players were convinced to go to Dove, relative to how many "mutant" *R* players (those playing Dove) it would take, starting from the "unnatural" convention of *R*-Hawk/*C*-Dove, to switch the *C* players towards their "natural" position of Hawk.

[27] McAdams, supra note 3, at 1673.

tial *formation* or *selection* of norms in games with unbalanced arrays—especially where these are based on "naturally salient" features. Our intuition is that oppressive norms can develop apace without centralized direction and that expressive efforts will have little or no effect on the adoption of these "natural" conventions.

IV. CHANGING NORMS

A. Alternative Stable Equilibria: Whose Ox Is Gored?

The analysis so far suggests that expressive law may have little to do with norms that emerge in unbalanced games where players fall into "naturally" salient categories and players indulge their "natural" tendencies for play. We would not expect law to have much influence on the focal feature around which the convention will tend to coordinate, nor would we expect much influence on the actual convention selected, which is a function primarily of player responses to expected payoffs.

This brings us to the issue of norm change. The challenge presented by the worst of oppressive norms is not one of incoordination and chaos but, if anything, too much uniformity and conformity in the "wrong" direction. So even if law as expression would appear to be unimportant to the emergence of oppressive conventions, there still remains the question of whether expressive law can be enlisted to help change existing norms. As already noted, this question matters a great deal: Better to use "cheap talk" than expensive (and imperfectly effective) penalties to effect changes in behavior.

The issue of how we might unseat oppressive norms, however, prompts us to take a fresh look at the question of why we might want to change them. In light of the analysis so far, this question is a potentially troublesome one. As we have seen, once a feature is chosen around which play can coordinate, there are only two "pure" strategy equilibria to choose from. In unbalanced games like those in the Appendix, the prediction is that a *C*-dominant/*R*-submissive convention will emerge "naturally." But what would it mean to "change" that convention? The only change that would establish a self-sustaining and stable convention—and also avoid the costly and undesirable "fourth box" of conflict—would be one that "turned the tables" on the players by flipping the pure-strategy re-

sult from the reigning *C*-Hawk/*R*-Dove to *R*-Hawk/*C*-Dove. Another alternative would be to try to move the game into the box of mutual cooperation, or the Dove/Dove combination. That strategy is inherently unstable, however: The parties have no incentive to adhere to it and continuous intervention to alter payoffs would be required to preserve the status quo. Although a spontaneous change in players' tastes could effect that result,[28] that state of affairs would not appear to be one that law could bring about, at least in the short term, without employing its sanctioning power.

This suggests that the changes in so-called oppressive conventions that the law could effect *expressively* will end up being a matter of table-turning or reassigning the roles of winners and losers rather than of mitigating the inequalities inherent in those roles. This highlights a hard truth about the social interactions modeled as Hawk-Dove games, including many coordinated around elements of players' "identity": Their very structure dictates that the interests of one class of players must be elevated over another as the price of fruitful cooperation and of minimizing costly social conflict.[29] Nothing short of revising payoffs would appear to abolish this constraint.[30] The point is that it is difficult to see how mere expressive intervention without continuous sanctions could produce and maintain any nonequilibrium outcome, at least over the long term. The model of expressive law that McAdams and others suggest is a clockwork conception that is largely self-executing: Law pushes the game in one direction but the parties' self-interest drives the mechanism.

[28] See discussion infra pp. 1754–55, 1760–63 of shifts in players' tastes and preferences as one mechanism of norm change.

[29] For a discussion about sexual harassment along these lines see, for example, Kingsley Browne, An Evolutionary Perspective on Sexual Harassment: Seeking Roots in Biology Rather than Ideology, 8 J. Contemp. Legal Issues 5, 37–39 & n.183 (Spring 1997) (asking "who is to blame" for psychological gender differences that lead to different perceptions of the meaning of workplace conduct, and suggesting there is no apparent reason why the law favors women's viewpoints or interests over men's); see also Marie T. Reilly, A Paradigm for Sexual Harassment: Toward the Optimal Level of Loss, 47 Vand. L. Rev. 427 (1994) (adopting a "neutral" or nonjudgmental stance towards the tradeoffs of men's and women's interests inherent in different sexual harassment rules or norms).

[30] See, e.g., Mahoney & Sanchirico, supra note 4 (using civil liability to alter payoff arrays and influence moves in the game by mandating transfers of resources between the parties).

Before proceeding to investigate how legal expression might change oppressive norms, we must inquire why we might want to effect a change that favors previous losers over current winners. One justification looks not to distributional effects but to efficiency. Some norms may be socially wasteful. If the sum total of payoffs to all players in the game is less than under some alternative stable convention, that alone may justify intervention to shift the norm. In contrast to games with similar payoffs for opponents,[31] different strategies for playing unbalanced games may generate quite different gains overall. As can be seen from the examples provided in the Appendix, the possible stable equilibria, including the convention that will emerge "naturally," will not always represent the most efficient combination of moves for the sum total of social welfare. In some cases, the disfavored equilibrium maximizes total utility. In others, the nonequilibrium cooperative combination (Dove/Dove) is most efficient.

As suggested earlier,[32] undesirable externalities or third-party effects provide another reason to try to "flip" the dominant convention. The smoker/nonsmoker game, which McAdams discusses at length, is one example. The costs and benefits to "players" reflected in the payoff values will not necessarily capture all consequences of a smoker-dominant convention. A society in which people are free to smoke will inevitably lure young people into addiction and inflict costs on third parties from the morbidity associated with smoking-related health problems and early death. These costs are left out of the idealized model of the game, which looks only to tastes and consequences for those on the front lines of the struggle to control public spaces. That narrow focus appears to dictate social indifference between smoker and nonsmoker dominance, since the interests of the two groups in "getting their way" appear to be equivalent and equally valid. Likewise, intrinsic payoffs to players from games that represent race and gender relations may not capture all secondary and downstream effects of structural inequalities that could flow from repetitive rounds of play over sustained periods. Lower payoffs to women within the

[31] Compare the discussion infra pp. 1755–56 on efficiency in games, with balanced arrays, suggesting that, absent externalities, there is no efficiency rationale for choosing one convention over another in this context.

[32] See discussion of the smoker/nonsmoker game, supra note 20.

division of labor or sexual harassment games might translate into reduced well-being for children or other undesirable social effects. Such consequences might justify interventions to discourage *C*- (or male-) dominant equilibria in some social situations.

Absent negative externalities or efficiency concerns, the struggle over which convention will prevail would appear to come down to rent seeking. Assuming the decision to favor one custom over another should rest on considerations other than raw favoritism, justifying legal intervention to turn the tables on prevailing norms would depend on finding some public-regarding basis for distinguishing between the possibilities represented by the available equilibria. Appeals to morality, responsibility, fairness, and right action could provide the grounds for regarding the status quo convention as less desirable than alternatives. Indeed, the lopsided structure and dynamics of "Chicken"-like social interactions makes it hard to avoid "value judgments" about the ways these games should be played. Those judgments may provide the only generally valid basis for deciding that the law should strive to establish new conventions that allocate benefits to some citizens at a cost to others. The decision to change the status quo must stem, in effect, from a determination that some types of interests or preferences should not be indulged but must be sacrificed in the name of what is morally right or socially just. An appeal to such impartial principles would seem to mark the only way out of the dilemma of partiality built into the structure of social interactions most aptly modeled by these games.

Sexual harassment (see Appendix) provides a good example of the law's attempt to work a sea change in conventions that previously favored men's interests in exacting sexual favors or dominating the workplace but now favor women's interests in avoiding those demands. The male interests sacrificed by the recent norm shift are now generally regarded as unworthy of official protection because they have come to be seen as harmful to many women's dignity and well-being.[33] The law's efforts to change the

[33] There is, of course, a range of conduct that potentially qualifies as sexual harassment under existing law, which is unsettled and ambiguous. Moreover, women's preferences (and thus their payoffs) in regard to outcomes and moves in the game are diverse, with some suffering intensely from conduct that others may not mind at all, and some (eligible single women, for example) potentially gaining in some cases from more lenient rules for mixing business and pleasure.

previously reigning convention stem from abandoning a stance of neutrality towards the parties' stakes in the game.

As noted, the law can attempt to disrupt the sexual harassment convention by altering players' actual payoffs through sanctions or penalties or by mandating transfer payments from wrongdoers to victims.[34] Alternatively, tastes may change for reasons that are ill-understood but may be partly influenced by changes in the law.[35] The limitations of the "deterrence" approach have already been discussed, and nowhere are those limitations more evident than for rules of interpersonal workplace conduct. The focus of my interest here, then, is on how law, apart from its power to sanction, might play a part in shifting the norms of conduct in the sexual harassment context and for similar interactions that produce disparate payoffs and coordinate around "salient" personal or identity traits. Before setting out a positive theory of how law might contribute to the reform of these kinds of "oppressive" norms, the next Section offers a more general critique of McAdams's analysis of how expressive law influences shifts in conventions and shows the limitations of applying the theory in this context.

B. Changing Conventions Through Expression: The Limitations of McAdams's Model

Part IV of McAdams's Article investigates how legal expression might work to change prevailing conventions for playing Hawk-Dove games. McAdams's analysis shows him struggling to come to terms with a central fact about these conventions: They represent stable—and thus *self-perpetuating*—strategies for the ongoing conduct of the game. These conventions are stable precisely because they are maintained by steps that players voluntarily and spontaneously take. The equilibrium is the result of individual players' decisions to make moves from which they have no good reason to deviate. The convention is difficult to dislodge precisely because no self-interested player can expect to gain by changing what he is doing. That is the fundamental hallmark of an equilibrium strategy.

Unlike some other games (such as the prisoner's dilemma), Hawk-Dove can fall into more than one stable conflict-free con-

[34] See Mahoney & Sanchirico, supra note 4.
[35] See text accompanying note 48, infra.

vention (corresponding to a "pure" Nash equilibrium). The payoff structure dictates that different players prefer different conventions. Although those who do worse under the status quo convention ("the underdogs") would prefer to switch to the more favorable regime, they won't do what is necessary to effect that change. By definition, no "rational" player—regardless of whether he could do better under a different convention or not—will choose to deviate from the existing mode of play. That deviation, if undertaken unilaterally, will be costly relative to sticking with the moves dictated by the existing norm. Consider Game #1 in the Appendix (p. 1777) played according to the "natural" convention of *C*-Hawk/*R*-Dove. *R* would prefer to live under the opposite convention of *R*-Hawk/*C*-Dove, which would enable him to more than double his payoff, from 5 to 12. *R* cannot get where he wants to go, however, without paying a steep price. Because he lacks the power to force *C* to do his bidding directly, *R*'s options are limited: He can cease playing Dove and start playing Hawk. But if *C* makes no corresponding change—and there is no objective reason why he should—*R*'s move entails sacrificing the certainty of a gain of 5 while incurring a loss of -10. *R* will not ordinarily choose to do that, so matters will remain as they are.

This analysis points to a feature that inevitably confronts any theory purporting to explain norm changes, where those changes are not centrally mandated and so must be effected by players' *voluntary* choices: All routes from one equilibrium to another are booby-trapped with losses for any player who seeks to initiate any reform. The basic puzzle for any account of informal norm change is thus to explain how, apart from directly altering payoffs to "force" players to act differently, parties can be persuaded to make the moves necessary to bring about the shift. Because the payoff array dictates that any isolated, unilateral decision to deviate from the convention will be penalized, it is hard to see how conventions can be changed by the players' self-initiated, uncoerced decisions. Any moves required to accomplish a switch between one equilibrium and another would always appear to be *against each player's rational self-interest.*

McAdams's effort to show how legal expression, operating noncoercively, might get around this problem is less than fully persuasive. His discussion centers almost exclusively on the dra-

matic recent shift from a smoker-dominant to a nonsmoker-dominant convention in public places. Central to his tale—and his story of how law figured in it—are two key elements: the potential for spatial segregation of smokers from nonsmokers and the existence of "cranks" among nonsmokers.

In keeping with the forgoing discussion, McAdams cannot help but recognize that, absent coercion of smokers by punishing them directly, nonsmokers' voluntary one-sided decision to defy the existing smoker-dominant convention is an unavoidable initial step in any sequence leading from smoker dominance to its opposite. Notwithstanding the announcement of an official smoking ban, the preexisting equilibrium erects a seemingly insurmountable barrier to playing by the new rule. Even if nonsmokers take the new rule seriously, there is no reason to believe that this step will be accompanied by any spontaneous change in smokers' or nonsmokers' conduct. As McAdams puts it, "The day before the state bans smoking in designated areas, 100% of smokers played Hawk in these (and all other) areas. The day after the ban, unless the law threatens sanctions, why would anyone expect anything different from smokers?"[36] If, as predicted, smokers continue to play Hawk even in nonsmoking areas, "then it does not pay for nonsmokers to challenge them."[37] The law's mere declaration of a new (unenforced) rule does nothing to budge players from their current course, because ignoring the rule continues to comport with every player's best interests.

McAdams sees the potential for spatial segregation as critical to moving out of this rut. For nonsmokers, the downside of switching from Dove to Hawk—the critical first step on the road to a new nonsmoker-dominant convention—is the possibility of encountering a Hawk in the form of an assertive smoker. One way to minimize this costly possibility is to arrange things so that nonsmokers encounter mostly their own kind. This will occur if nonsmokers flock together. The official designation of "nonsmoking" areas facilitates segregation by allowing nonsmokers easily to find one another. If nonsmokers can be assured of encountering mostly nonsmokers, the probability of getting into a fight will de-

[36] McAdams, supra note 3, at 1717.
[37] Id.

cline precipitously, and with it the costs of standing ready to play Hawk. In effect, spatial segregation helps cushion the cost to nonsmokers of defying the existing smoker-dominant convention by ensuring that nonsmoker's willingness to go up against smokers is tested very infrequently.

Unfortunately, however, segregation is a two-way street. This steep decline in nonsmokers' costs of assertiveness will occur only if Hawkish smokers also stay away. If smokers go everywhere, nonsmokers will receive less than full protection from the costs of belligerence. But there is no obvious reason why smokers won't go wherever they like. Because nonsmokers can only afford to be bold if their boldness is never or rarely tested, nonsmokers will not shift their strategy if smokers invade their turf in large numbers. All "rational" nonsmokers will back down when confronted with aggressive smokers, so the latter have no reason to stay away or to shrink from their customary aggressiveness.

This is where "cranks" come into the picture. Cranks are defined as those outliers who "will play [hawk] regardless of what they expect the other to do."[38] For cranks, "the game being played is *not* Hawk-Dove" because "Hawk is the *dominant* strategy," either because the gains from winning a fight are much higher than for other players or the costs of losing much lower.[39] Although cranks, like all other players, still prefer their opponents to behave submissively, they will stand up to assertive behavior rather than yield to it. Cranks are not afraid to fight.[40]

According to McAdams, nonsmoking cranks are the solution to the problem of smoking Hawks in the nonsmoking section. Cranks discourage Hawks from invading nonsmoking areas because the possibility of encountering a crank, even if fairly low, is sufficiently unpleasant to induce almost all Hawkish smokers to stay away. McAdams insists that cranks need not be numerous to play their role effectively.[41] Because cranks provide smokers with a reason to

[38] Id. at 1718 n.142.

[39] Id.

[40] For another example of crank-like behavior, see Posner's description of the willingness of the destitute to flout the norms of property ownership. See Posner, supra note 4, at 30. Notoriously, the poor may resort to criminal activity if their stake in existing rules evaporates because they have no property to call their own. See id.

[41] According to McAdams, nonsmoking cranks will also tend to stay in the nonsmoking section despite their low costs (or net gains) from encountering smokers

avoid nonsmoking areas, smokers will begin to confine themselves to other areas. The two-way process of segregation proceeds apace.

The last step in McAdams's story sees this bilateral self-segregation as fostering a change in expectations. The resulting emergence of smoke-free spaces gives smokers and nonsmokers alike "significant reason to wonder whether past precedent predicts future play"—that is, whether players of all stripes will continue to adhere to the smoker-dominant convention.[42] In effect, smokers begin to wonder whether nonsmokers will start to resist and nonsmokers wonder whether smokers will continue to do so. As with McAdams's account of how norms evolve in the first place, timing is important: Expectations and actions must change in concert for players in both camps. Smokers and nonsmokers must be equally willing to switch their strategies when encountering their opposite number. If most nonsmokers begin playing Hawk and most smokers play Dove, that will go a long way towards shifting the equilibrium. The more players on both sides play the game the new way, the more the remaining players do best by following suit.

The plausibility of McAdams's account would appear to depend on the validity of myriad assumptions that must be evaluated empirically. Whether a few well-placed cranks can succeed in inducing most smokers to avoid no-smoking areas would seem to depend on the size, attractiveness, location, and convenience of alternative spaces and other factors varying case by case. In addition, McAdams asks us to accept that the mere fact of segregation will cause players to believe their opponents will abandon the old convention in considerable numbers and play by the new legally ordained rule. It is hard to see why this should occur and impossible to know *a priori* whether it actually will.

These difficulties are not central to this critique, however. Rather, this account identifies as the most important flaws in McAdams's analysis its reliance on two key features: spatial segregation and the presence of "cranks." To the extent that McAdams's account of convention change turns on self-segregation, its potential application to other contexts and other games is severely limited. Spatial or social segregation is not a feasible prospect for many in-

because their costs are minimized even more by avoiding them. See McAdams, supra note 3, at 1719 n.142.

[42] Id. 1719–20.

teractions that are of the greatest societal interest. In many cases segregation would defeat much of the purpose of reforming prevailing norms. Spatial or social segregation by race or sex is often regarded as an independent source of subordination that is undesirable in itself. Even if temporary, segregation would often greatly reduce the benefits for the disfavored group of playing the game at all.[43]

Second, reliance on the presence of cranks provides a tenuous basis for a generalizable theory of norm change. Although, as McAdams correctly states, "individuals vary continuously and some have idiosyncratic payoffs that produce no incentive to ever follow the convention,"[44] cranks necessarily represent an atypical extreme—so much so that Hawk is a dominant strategy for them, which negates a fundamental assumption of the game. Because "crankiness" betokens a wide divergence from most players' preferences, it is safe to predict that cranks will be rare indeed among "underdogs" in most settings that give rise to "oppressive" norms. Unless cranks need only be present in small numbers, any account of norm shifts that assigns them a key role must be viewed with skepticism.

The most important objection to reliance on cranks, however, is that societies need not accept cranky behavior passively. Those who fail to avoid conflict spontaneously represent a serious threat to social order and may be targeted for special measures of social control; these may include joint third-party sanctions such as disapproval, ostracism, private violence, or collective economic pressure.[45] These measures would operate like official penalties to

[43] Spatial and social segregation of the sexes can be partial or can involve complex arrangements compatible with some cooperative interactions. And voluntary racial self-segregation is favored by some as a solution to the dilemmas of racial subordination. As a general matter, however, segregation of groups would be considered too controversial, too difficult to effectuate, or too costly to employ as a method for changing undesirable conventions.

[44] McAdams, supra note 3, at 1696 n.108.

[45] These social pressures could arguably be regarded as affecting outcomes for players in ways that go beyond those captured directly by the payoff array. A woman who complains about sexual harassment may be ostracized by family or friends, which adds to the costs incurred from her opponent's retaliatory moves. A white restaurateur, although personally averse to serving blacks, may hesitate even more from fear of ostracism or recrimination. There are other examples in which norms are reinforced by outside pressure. The source of the collective impulse to shore up conventions is not well-understood; it may represent a response to the threat to the stability of norms created by the diversity of player preferences.

alter cranks' payoff array and enhance the costs of assertiveness, thus ensuring that moves reflect not just the players' preferences but the interests of the community's dominant groups. By selectively ratcheting up sanctions for troublemakers, groups can "reinvert" the order of preferences for the outliers who threaten the stability of established conventions. That cranks can be and often are effectively "neutralized" by such devices undermines the notion that cranky behavior will be sufficiently robust to fuel observed shifts in conventions.

In any event, the manner in which cranks function in McAdams's smoking example is necessarily of limited interest because their role is not independent of the potential for spatial self-segregation. In McAdams's account, spatial segregation and cranks are factors that work in concert to bring about the smoker convention change, and each component is necessary to the story. Thus even if cranks are not uncommon in analogous situations, McAdams's account is too parochial to explain much norm change.

The case of sexual harassment (see Appendix, p. 1779) illustrates the limited generalizability of McAdams's analysis. The problem of harassment arises only because men and women are thrown together at work. Sex segregation in the workplace would certainly "solve" the problem of harassment by reducing the occasions for conflictual encounters. But even if temporary segregation were a price worth paying for an enduring change in norms, it is difficult to construct an account analogous to McAdams's smoking story for how sex segregation plus cranks might hasten the shift from the male-dominant status quo to a harassment-free environment.[46]

A role for cranks in changing sexual harassment norms is problematic for two reasons. First, as predicted, they will be rare. The payoffs for female cranks would have to diverge quite dramatically from those typical for most women, with the expected benefits of resisting and/or complaining far exceeding the downside risk of going head to head with a vengeful and insistent male harasser. This could happen, as McAdams suggests, if a woman's gains from "getting her way" in a fight exceeded her losses from a rout. Since the

[46] Even if such an account could be devised, many feminists would insist that transitional segregation would carry intolerable costs that render it an undesirable device for helping resolve conflicting male and female priorities at work. See discussion supra note 43.

outcome of any given fight is not predetermined, however, the Hawk/Hawk payoffs supplied in game theoretic arrays generally represent the *expected* costs of coming into conflict with an opponent. This in turn depends on the probability of a positive or negative outcome as well as the utilities derived from each state. McAdams states that the payoffs supplied in his arrays are based on the assumption "that a player who 'fights' wins half the time."[47] That scenario may be unrealistic in the world of sexual harassment, however. If men win fights *most* of the time—a not implausible assumption—fighting would only be worthwhile if the underlying benefits to a woman of winning a fight exceeded the costs of losing many times over. Although occasional women may be so attached to their virtue or unattached to their jobs that this ratio obtains, such women will likely be few and far between.

But even if those women do exist, it is still not obvious how these isolated outliers could turn the tide on the entrenched male-dominant convention. As long as cranks are rare, men will encounter them quite infrequently and it will remain in most men's interest to stick with the established convention and play Hawk. Likewise, because the probability of encountering a Hawklike male would still be exceedingly high, it would remain in the ordinary woman's interests to stick with Dove. In McAdams's example, cranks' rebellious behavior works by enhancing self-segregation: Staying away is an easy and, presumably, low-cost way for smokers to avoid costly cranks while still continuing to smoke. Spatial segregation then leads to the expectation that nonsmokers will convert to Hawk. Even if this scenario is plausible (which it may not be), it is hard to see how a parallel sequence would operate in the harassment context. Even if women surrounded themselves with women and the female population was peppered with a few cranks, it is not obvious that men would consider it in their interest to stay away since—unlike with smoking—self-segregation would prevent them from having their cake (avoiding nasty conflict) and eating it too (enjoying the benefits of Hawklike harassment, which requires access to women). Even if they did stay away (and the employer let them), it is not obvious how separation of the sexes would operate

[47] McAdams, supra note 3, at 1719 n.142.

to generate analogous expectations of women's future aggressiveness in would-be harassers' minds.

V. NORM CHANGE AND EXPRESSIVE LAW: AN ALTERNATIVE ACCOUNT

This Section attempts to improve on McAdams's story by putting forward an alternative account of how changes in informal norms can occur. The analysis dispenses with reliance on the two problematic elements—the potential for segregation and the presence of cranks—identified above. The alternative theory put forth here is especially concerned with supplying a more plausible explanation for recent historical shifts in "oppressive" norms—that is, conventions that arise in unbalanced games coordinated around fairly rigid roles.

Although McAdams's analysis seems inadequate to the task of explaining how legal expression can help dislodge established conventions in Hawk-Dove type interactions in general—and in those giving rise to oppressive conventions in particular—the fact remains that such conventions do sometimes undergo dramatic reversals. Relations between racial groups and the sexes have been areas of active social ferment and our intuition is that direct enforcement of legal mandates has played at most an ancillary role in effecting these developments. Changes in tastes and preferences, with corresponding shifts in payoff values, have almost certainly had an important influence, although the processes that have fueled these changes are themselves somewhat mysterious and cry out for further explanation.[48] Although McAdams's failure to find a credible place for legal expression in explaining norm shifts is discouraging, we should not so readily give up on a role for expressive law in fomenting social change in these areas. The suggestion here is that law in its expressive capacity has figured to some extent in these movements for social reform, but that McAdams's analysis does not capture what is really going on when norms begin to favor the previously disfavored.

The account presented here, although not limited in application to "oppressive" norms, nonetheless has special power in explaining changes in these types of conventions because it taps into the po-

[48] See discussions supra notes 25 and 32.

litical psychology of those who live and suffer under customs that perpetuate persistent "identity-based" hierarchies. That psychology makes those conventions particularly vulnerable to alteration by the mechanisms described.

The alternative account proceeds from the assumption that real human beings often deviate from the strictly rational behavior that classic game theory takes for granted. That the rational actor model often fails to explain human decisionmaking is a persistent theme of recent work in behavioral economics.[49] Our analysis of McAdams's account of norm change reveals that the assumption of players' strict adherence to rational self-interest makes it difficult to explain how conventions representing stable game-theoretic equilibria *ever* come to change. By definition it is in no player's interest to defy them or to take the steps necessary to change them. But this raises the possibility that perhaps norm changes occur because people sometimes act *against* rational self-interest.

How could acting against narrow self-interest induce a change in the normative strategies for playing the games of interest here? Extending insights found in the work of Edna Ullman-Margalit suggests a possible answer.[50] Starting from the commonplace observation that stable conventions in many games yield unequal payoffs to opposing players, Ullman-Margalit examines examples of games that have settled into what she describes as a "status quo of inequality which is in a game-theoretical equilibrium."[51] She designates the conventions for playing these games "norms of partiality."[52] She suggests that many norms of partiality, despite their game-theoretic stability, are potentially vulnerable to displacement in real life,[53] and she describes such conventions as "strategically unstable."[54] Although recognizing that such conven-

[49] See Behavioral Law and Economics (Cass R. Sunstein ed., 2000) (summarizing recent work reaching this conclusion).

[50] See Edna Ullman-Margalit, The Emergence of Norms (1977).

[51] Id. at 163.

[52] Id. at 134.

[53] See id. at 162–63. As discussed above and in other sources, game-theoretically stable conventions are vulnerable to destabilization and displacement through the operation of "rational" evolutionary mechanisms in combination with mutations, "mistakes," and random noise. See, e.g., Bendor & Swistak, supra note 26 (summarizing theories of the robustness of equilibria). See other sources, supra note 26. Without denying those mechanisms, Ullman-Margalit does not make them the focus of her analysis.

[54] Ullman-Margalit, supra note 50, at 163 (emphasis omitted).

tions satisfy the strict Nash equilibrium condition that no player, including the "underdog" disfavored by the convention, has a rational incentive to deviate from the convention of play, she identifies the instability of such conventions as stemming from some players' *irrational* willingness to deviate from them. That is, a player may elect to defy the convention even if doing so is against his own self-interest.

Why would he do that? Ullman-Margalit's discussion suggests a distinction, important for our purposes, between the psychological *motivation* behind such a move and the strategic advantage that may eventually be gained from it. To illustrate, consider the following array:

		B	
		C1	*C2*
A	*R1*	2, 1	0, 0
	R2	0, 0	1, 2

Suppose the status quo convention for playing this game is (*R1*, *C1*). This is an equilibrium strategy, from which no player has reason to deviate so long as others adhere to it. Nonetheless, *B* might intensely dislike that equilibrium. One reason he might dislike it, Ullman-Margalit suggests, is that he assigns an independent value to improving his position *relative* to his opponent. Or he may value equality for its own sake. He may feel so strongly that he is willing to move unilaterally towards a more equal combination of payoffs even at the cost of forgoing benefits or incurring costs as reflected in the values in the array. In the game above, *B* might decide to switch from *C1* (with payoff 1) to *C2* (with payoff 0). Assuming that *A* will at first continue to play as before (that is, to adhere to the *R1* strategy), that move will succeed in effecting *B*'s immediate goal of making payoffs to the parties more equal and improving *B*'s welfare relative to *A*'s.

Although it can be debated whether the "preferences" that drive *B*'s moves under this account simply represent an alternative form of rational choice or a deviation from "true rationality," that dispute is unimportant for our purposes. Rather, as Ullman-Margalit suggests, this example relaxes the common game-theoretic assumption that each participant rank-orders his own outcomes *"in isolation* from the interaction situation, or in a sense even *prior* to

it."[55] The example here permits players to look to the overall pattern of payoffs in assigning a rank order to states of affairs and particular moves. Those concerns may dictate decisions taken in defiance of immediate "rational" self-interest, as reflected in the utilities assigned to each party in the array. What matters is the possibility that people may look beyond what is in the game for them alone; they may be motivated by something other than a narrow calculus of self-interest.

The "motives" that impel *B*'s decision to move from *C1* to *C2* are not the end of the story, because *B*'s move presents the potential for strategic advantage more narrowly defined. If *A* becomes convinced that *B* prefers "equality in misery" to the preexisting inequality despite the former being worse for *B*, A is faced with a dilemma. He can make do with his lower payoff under *R1*/*C2*. Or he can try to make the best of the situation by moving to *R2*, which has the effect of raising payoffs for both players. But that move also has the effect of reintroducing inequality, where that inequality now favors *B*. This sequence reveals that by doing something that is "irrationally" self-defeating in the short term, *B* has succeeded in bringing about a "new and advantageous . . . status quo" in which the tables of partiality are turned, but this time to his benefit.[56] *B* has managed to effect this reversal, paradoxically, only by persuading his opponent that he is "so indignant at the [unequal] status quo that he actually prefers anything, including an 'equality in misery' to it."[57]

A number of features in Ullman-Margalit's discussion point the way toward an account of norm change for the games that are the focus of this Comment—an account that offers a potential entry point for the influence of expressive law. First, Ullman-Margalit's analysis has the virtue of confronting head-on the coordination problem that confounds any self-executing shift between stable conventions. As Ullman-Margalit suggests, it is clear that no rational and self-interested player in the games she analyzes will unilaterally take the steps necessary to convert to a new equilibrium even if that equilibrium would ultimately prove better for him. That player will not move without a simultaneous change on

[55] Id. at 146.
[56] Id. at 167–68 (emphasis omitted).
[57] Id. at 167 (emphasis omitted).

the part of his opponent. A norm change requires that *both* parties deviate from the status quo. But a simultaneous change is not in the cards because "there is every reason to expect [the opponent] not to cooperate."[58] Ullman-Margalit recognizes that the solution to this dilemma is to be found in identifying "some method of inducing the other party to choose in one's favour."[59] Moreover, "[t]his inducement will be achieved once one succeeds in affecting the other party's expectations concerning one's own behaviour."[60] She suggests that a player "will be able to influence the other's expectations" only through "visibly and persuasively constraining [his] own behaviour."[61] In her example, *B* must somehow succeed in convincing *A* "that he is committed, in a binding and irrevocable way, to abandon the status quo."[62] *B* must indicate his readiness to make "an unconditional choice of column 2," thereby turning "an originally four-state situation . . . to only a two-state situation."[63] The challenge for *B*, then, boils down to making a credible case that he is firmly committed to a move *against self-interest.*

How might a player accomplish this? The work of economist Robert Frank is suggestive.[64] Frank describes circumstances in which individuals might wish to make "a binding commitment to behave in a way" that is "contrary to self-interest."[65] Players in the prisoner's dilemma game would be better off if both decided to cooperate. Yet defection is the only "rational" strategy for all players, and defect/defect the only stable equilibrium. Playing the game in a way that benefits all players simultaneously therefore requires players to make choices that depart from immediate self-interest. What is needed is a device that commits people to mutual cooperation, but without making them vulnerable to exploitation. Frank's idea is that the emotions—and specifically the moral sentiments—serve as that device in many real-life settings. Players' interest in fairness and reciprocity, as well as their intrinsic sense of

[58] Id. at 166.
[59] Id. at 165.
[60] Id.
[61] Id.
[62] Id. at 166 (emphasis omitted).
[63] Id. at 165.
[64] See Robert H. Frank, Passions Within Reason: The Strategic Role of the Emotions (1988).
[65] Id. at 47.

gratitude and indignation towards those who aid or betray them, impel them to cooperate with others, but only when other persons show willingness to cooperate with them. The sentiments that compel players to engage in conditional cooperation are also evident to observers and help communicate a *commitment* to behave in this way. The moral sentiments thus play the dual role of actually ensuring adherence to actions that are not strictly rational, and also of signaling to others the serious intention to take those actions. Frank implies that the best way to persuade others of one's firm determination to act against interest is through devices that are well known to impel such actions. Emotions familiar to all, if overtly displayed, operate as a commitment device that is both reliable and credible.

Frank's ideas about the role of the moral sentiments have potential application to the social interactions under consideration here. Frank's theory recognizes that although acting against rational self-interest may appear costly in the short run, it can yield benefits eventually by facilitating fruitful cooperation.[66] In our case, moves that are costly in the present can produce gains in the future. In both cases, emotions play a key role in motivating the actions against interest. Sentiments such as envy, outrage, indignation, gratitude, vengefulness, or a desire for fair treatment both impel the self-defeating course of action and signal a party's commitment to it.

Ullman-Margalit's discussion of the strategic instability of "norms of partiality," although underdeveloped, has much in common with Frank's insights about the role of emotions in signaling commitment and achieving coordination. In discussing her example of a norm of partiality, she suggests that "underdog" group members can only realize their desired goal of changing to a more favorable convention by persuading "top dog" opponents of their readiness to unilaterally abandon the current strategy despite the losses this entails. If the top dogs believe that underdogs are fully committed to acting against interest come what may, they will do well to alter their strategy as well. Because, as Ullman-Margalit recognizes, the underdog's commitment to defying the convention

[66] See, e.g., id. at 51–52 (drawing the distinction between immediate psychological motive and ultimate strategic effect).

will effectively reduce a "four box" option to two, the opponents' best choice will be to abandon an aggressive posture and adopt a more submissive role. Ullman-Margalit identifies the moral sentiment of indignation as playing a key role in effecting this progression. *B*'s contemplation of his inferior relative position or the unequal state of affairs fuels his resentment. His indignation prompts him to abandon the position that perpetuates this obnoxious situation even at great cost to himself. The top dogs are acutely aware that the underdogs feel resentment and are likely to act on their feelings in a predictable way.

Although Ullman-Margalit does not unpack the sequence carefully, her discussion suggests that the underdog's outrage plays the dual role suggested above: It impels that party's shift in strategy, and it provides the basis for the opponents' belief that the underdog will stick to his course despite personal sacrifice. The emotions felt by the previously submissive group also function "horizontally" to foster group solidarity. Such solidarity serves to facilitate coordinated action and to speed change. If some underdogs advertise their commitment to defeating the status quo, others will be emboldened and encouraged to follow. The greater the number of underdogs joining the movement for change, the more costly it is for members of the previously dominant group to continue on their current course. The expectation that underdogs will act en masse in defiance of the existing convention will tend to weaken opponents' commitment to it by raising the intolerable specter of conflict.

Although she does not label them as such, what could be termed "moral sentiments" or "moral attitudes" are critical to Ullman-Margalit's account of how players unsettle game-theoretically stable equilibria.[67] She ascribes the underdog's willingness to disrupt the status quo to his desire for equality or improved relative position and acknowledges that strong human emotions drive these desires. She also recognizes, in keeping with Frank's analysis, that the ultimate result of putting these emotions to work is strategic. The end-point of the disruption is not greater equality of resources overall but a convention that effectively reverses the positions of the players.

[67] See, e.g., Ullman-Margalit, supra note 50, at 155, 165–168.

Although this analysis suggests that the scenario should be regarded as a mere battle of wills or a power struggle that rewards the party willing to sustain the heaviest losses, those are not the terms on which movements for social change actually proceed in real life. Rather, the push for reform is routinely informed by recourse to moral conceptions of right and wrong, justice and injustice. The thrust of the rhetoric ordinarily employed by critics of prevailing social practices in realms of sex and race is that certain methods for gaining advantage through private dealings are morally wrong, deserve condemnation, and should be off-limits. People are not supposed to use their position at work to gain sexual favors. One should not satisfy one's personal preferences by excluding blacks from privately owned public accommodations. Husbands are not supposed to gain the upper hand over their wives through superior bargaining power. And so on.

These insights point to the pivotal importance of the *moral* element in reactions to the norms that govern social life. As already suggested,[68] it is important to any movement for social change to persuade others that not all conventions are equally desirable. But desirable to whom? Because it provides a requisite principle of impartiality that allows proponents to rise above the appearance of seeking advantage at others' expense, moral argument is a key element in the battle to change norms. This suggests that social movements cannot entirely dispense with moralistic judgments.

The moralistic character of players' motives and justifications plays both a psychological and a rhetorical role; both are vital to players' perceptions and behavior within the game. The emotions the status quo elicits in the disfavored class stem from the perception of being treated wrongly or unjustly. Resentment is directed not at losing out as such but at *wrongfully* imposed disadvantage. That the particular arrangement at issue is not only unequal but also unfair helps resolve the tension between the motivating sentiments and the "table-turning" final outcome.

The rejection of an exclusive reliance on self-interest in favor of an abstract concern with justice and right also confers an important strategic advantage, as it enhances the credibility of the underdog's commitment to act "irrationally." If the underdog's motive is a

[68] See supra pp. 1763–66.

mere selfish desire to change places with favored players, he might abandon his strategy if the costs of adhering to it grow too great. The perception that a desire to gain advantage is at the root of the push for social change invites an opponent to ratchet up the costs of defying convention, as high costs can be expected to weaken resolve. It makes more sense to act against self-interest, however, if the objective is not to gratify the self but to vindicate principle. But the devotion to principle cannot be shaken by imposing losses, if avoiding these losses is not the reason for commitment. A moral rationale thus provides a far more credible platform for convincing the opponent of one's willingness to endure whatever costs may be imposed. By suggesting that draconian countermoves won't work, a moral justification discourages escalation and encourages the conclusion that there is no choice but to give in to the desired change. In sum, the commitment to broader moral principles, and the link to the sentiments triggered by commitment to those principles, may well prove essential to the effectiveness of many efforts to reform "norms of partiality."

Assuming this account is plausible, how might legal expression facilitate or shape this process? In venturing a guess, it is important to note that the picture suggested here relies on a richer model of human motivation and behavior than McAdams sees the need to employ in his account. It assigns a place to moral sentiments and emotions elicited by wrongful or unjust treatment, and it accepts that people will sometimes act unselfishly from morally inspired motives. Understanding how law might influence this dynamic may require recognizing an expressive power for law that goes beyond imparting information, focusing citizens' attention, or fostering expectations about behavior. Experience suggests that law might possess a hortatory or normative power—that is, it might possess a capacity to influence what citizens regard as morally justified or right. Although as yet poorly understood, this aspect of law may have something to do with law's ability to influence the moral struggle over convention reform.

By announcing rules that run contrary to existing informal conventions (as with laws against sexual harassment or race discrimination), the law may add normative weight to outcries by private citizens, groups, or organizations against existing arrangements. The law's imprimatur may embolden and encourage underdogs to

take personal risks in defiance of existing conventions. Any presumption, however mild, that law and morality coincide fuels the underdogs' conviction that they are justified in acting against interest to vindicate higher values and reinforces the opposing groups' expectation that underdogs will continue to defy the status quo. The desire to avoid costly conflict will be the opposition's principal reason for its willingness to switch strategies, but the highly moralized character of the struggle will also help undermine the opponents' determination to continue on the present course.

The law's publicity function can also facilitate concerted effort by calling attention to the possibility of a different convention and by putting a seal of approval on a state of affairs that runs contrary to the status quo. By publicly advertising and affirming the "rightness" of the new convention, the law can raise awareness of the injustice of the existing regime, add momentum to private resentments, and provide a focal point around which players can rally and affirm their commitment to change. By thus encouraging and advertising rebels' strength, sincerity, and determination, the law helps weaken opposition to reform. Although effective sanctions will certainly speed this process along, this account suggests that the communicative power of law can have some effect as well.

The alternative theory sketched here improves on McAdams's account in a number of ways. One advantage is that it does not assign a crucial role to cranks. Norm changes can be spearheaded by players who, although holding a range of preferences, do not differ much from the average player in the costs they incur from conflict. It is not the lower expected costs of conflict, but the countervailing emotional, psychological, or moral commitments to change that make underdogs willing to face the risks of a fight. They will fight even though they know that they could well *lose*, at least on the initial rounds. Whether, as noted, this is best conceptualized as an effective "crank-like" modification of players' payoffs or an example of unselfish behavior[69] is irrelevant for our purposes: The critical difference lies in providing a more authentic and accurate understanding of the *political psychology* of norm change. The central point is that those disfavored by the existing status quo must convince themselves to engage in self-sacrifice as the only route

[69] See supra pp. 1754–60.

towards a "better world." The main impetus for defiance is the indignation engendered by unfair and unjust treatment and a conviction that the status quo is morally wrong. Emotional, psychological, or moral commitments, not a strictly rational cost-benefit calculus, mediate the behaviors that will ultimately enshrine a new social order. Those commitments spur willingness to act despite considerable personal costs.

Another advantage of this account is that it is more consistent with historical and social reality. Moralistic rhetoric and appeals to justice are a pervasive hallmark of movements to reform or abolish oppressive norms. This story explains the observed patterns by showing how moral sentiments motivate the steps victims must take and also mold oppressors' expectations. Second, in assigning a key place to the willingness of underdogs to compromise immediate self-interest on behalf of their goals, the account is consistent with a strong role for self-sacrificial leadership and "martyrs to the cause"—a pattern that comports with the history of the twentieth-century civil rights and feminist movements. The willingness of high-profile leaders to incur personal costs serves as a coordination device that inspires the rank and file to similar sacrifice. And without such mass sacrifice there can be no norm change.

Finally, this account suggests an explanation for the paradox that the very conventions which by virtue of their "oppressiveness" appear particularly resistant to subversion may in fact prove especially vulnerable to destabilization. The examples of oppressive unbalanced games in the Appendix have in common that underdogs lose more than their opponents from conflict. The payoffs are such that players who gain least from the status quo also stand to suffer most from defying it. But the greater their self-interested costs of conflict, the greater the barrier to underdogs' unsettling the dominant convention by rebelling against it. This pattern would appear to pose formidable obstacles to reform. Indeed, Edna Ullman-Margalit herself despairs of the prospect of

changing game-theoretically stable conventions in which the route out entails significantly greater losses for one party than the other.[70]

This analysis shows why her pessimism may miss the mark. Recent history suggests that "oppressive" conventions that develop in situations with very lopsided payoffs—including those that pose significantly greater downside risks for disruptive underdogs—have not proved particularly stable, at least in the long term. The very fact that disfavored nonconformists potentially face egregious losses—and losses greater than opponents'—may both add to the underdogs' moralistic fervor and make that fervor more credible. The prospect of greater losses means greater self-sacrifice. But greater self-sacrifice signals greater moral seriousness. The more self-interest must be compromised to effect change, the more likely it is that moral commitment and not just self-interest is at work. But if moral commitment is the true motive, the more likely that the underdog will stand firm despite considerable personal costs. The degree of oppressiveness and the effectiveness of the commitment strategy thus play into one another, fueling the process of reform.

In addition to disparities in payoffs, the intransigence of role assignments that is the hallmark of the most oppressive norms may also contribute to those norms' instability by reinforcing the underdogs' resolve. It is harder to stir up moral indignation about temporary disadvantages that are voluntarily assumed than about those imposed by circumstance. The more inflexible the role assignments within the prevailing convention, the more entrenched and enduring are the disadvantages it creates. Greater injustice produces greater determination to change it.

---

[70] See Ullman-Margalit, supra note 50, at 161. In commenting on this matrix

		B	
		C1	*C2*
A	*R1*	10, 6	4,-30
	R2	5, 4	8, 8

Ullman-Margalit explains that

> the state represented by the top left R1-C1, which is neither equitable nor optimal for [*B*] (who would obviously have preferred state R2-C2) is nevertheless a game-theoretical equilibrium which is at the same time stable by our [strategic] standards as well: it is on no account susceptible to threats. [*A*] enjoys it, while [*B*] is intimidated from abandoning it by the fear of getting -30.

Id.

If we accept that law can operate expressively to speed and strengthen these developments, then this analysis suggests that expressive law might be particularly effective in helping change the *most* oppressive norms. The implication is that even where customs and interests are most entrenched and views most polarized—which is also where enforcement may prove most difficult, costly, or imperfect—changes in law might be well worth pursuing. The history of the civil rights and feminist movements bear out these intuitions. Many of the laws enacted to effect change in these areas have been enforced imperfectly or only with great difficulty. This analysis suggests they may have been worth enacting despite these limitations.

CONCLUSION

The capacity of governments to deter conduct by directly threatening and imposing sanctions is limited by burdensome costs of enforcement and imperfect detection, prosecution, and proof. Drawing on law in its expressive capacity is a potentially less onerous way to bring about desirable social change. Because governments should opt in favor of employing the expressive power of law over its deterrent power, it is imperative to understand how law can produce change expressively if possible. This Comment offers one account of how entrenched private norms of behavior can change despite their seeming stability, and how legal rules can influence that change. It suggests that laws against oppressive private behavior may at times be worth enacting despite formidable obstacles to conventional modes of enforcement.

APPENDIX

Game #1

		C	
		Dove	*Hawk*
R	*Dove*	8, 5	5, 15
	Hawk	-12, 2	-10,-2

	Cooperation	*Conflict*	*R dominant*	*C dominant*
Total utility (R + C)	**13**	-12	14	**20**
R utility	8	-10	12	5
C utility	5	-2	2	15

Expected payoffs given a 50% chance opponent will play Hawk/Dove:

R from Dove = 8 + 5 = 13
R from Hawk = 12 - 10 = 2

C from Dove = 5 + 2 = 7
C from Hawk =15 - 2 = 13

Game #2

		C	
		Dove	*Hawk*
R	*Dove*	8, 6	0, 11
	Hawk	10,-2	-10,-5

	Cooperation	*Conflict*	*R dominant*	*C dominant*
Total utility (R + C)	**14**	-15	8	**11**
R utility	8	-10	10	0
C utility	6	-5	-2	11

Expected payoffs given a 50% chance opponent will play Hawk/Dove:

R from Dove = 8 + 0 = 8
R from Hawk = 10 - 10 = 0

C from Dove = 6 - 2 = 4
C from Hawk =11 - 5 = 6

Sexual Harassment Game

Dove for men is to refrain from harassment/retaliation.
Dove for women is to yield (give in, tolerate, or leave job).

Hawk for men is to harass and/or retaliate.
Hawk for women is to resist (refuse and/or complain).

* * *

Dove/Dove outcome is compromise (for example, each takes turns playing Hawk to the other's Dove, or getting his/her way half of the time).

Hawk/Dove (male/female) is for man to harass, woman to yield.

Hawk/Dove (female/male) is for women to (stand ready) to resist/complain, men to refrain from harassment/retaliation.

Hawk/Hawk is for men to harass/retaliate, women to resist/complain.

* * *

C (male)

R (female)		*Dove*	*Hawk*
	Dove	0, 5	-3, 12
	Hawk	13,-1	-20,-5

	"Cooperation"	*Conflict*	*R dominant*	*C dominant*
Total utility (R + C)	5	-25	12	9
R utility	0	-20	13	-3
C utility	5	-5	-1	12

Expected payoffs given 50% chance opponent will play Hawk/Dove:

R from Dove = 0 - 3 = -3
R from Hawk = 13 - 20 = -7

C from Dove = 5 - 1 = 4
C from Hawk = 12 - 5 = 7

Division of Labor Game

Hawk is to do paid but no unpaid work (for example, domestic labor).

Dove is to do unpaid work (and, or instead of, paid work).

		C (male)	
		Dove	*Hawk*
R (female)	*Dove*	8, 6	-3, 13
	Hawk	11, 2	-3, 0

	"Cooperation"	*Conflict*	*R dominant*	*C dominant*
Total utility (R + C)	14	-3	13	16
R utility	8	-3	11	3
C utility	6	0	2	13

Expected payoffs given 50% chance opponent will play Hawk/Dove:

R from Dove = 8 + 3 = 11
R from Hawk = 11 - 3 = 8

C from Dove = 6 + 2 = 8
C from Hawk = 13 + 0 = 13

Public Accomodation Game

Assume that segregation is neither mandated nor illegal, but that majority group customers shun integration.

Hawk for *R* is to insist upon accommodation.
Hawk for *C* is to refuse accommodation.

Dove for *R* is to avoid majority-owned establishments.
Dove for *C* is to accommodate minority group customers.

Assume Dove/Dove is a "split the difference" compromise: *R* avoids half the time, *C* accommodates half the time.

C (majority)

R (minority)		*Dove*	*Hawk*
	Dove	7, 6	0, 12
	Hawk	10,-2	0,4

	"Cooperation"	*Conflict*	*R dominant*	*C dominant*
Total utility (R + C)	13	-14	8	12
R utility	7	-10	10	0
C utility	6	-4	-2	12

Expected payoffs given 50% chance opponent will play Hawk/Dove:

R from Dove = 7 + 0 = 7
R from Hawk = 10 - 10 = 0

C from Dove = 6 - 2 = 4
C from Hawk = 12 - 4 = 8