ARTICLES

AN EVOLUTIONARY APPROACH TO NORMS

ROBERT AXELROD
University of Michigan

Norms provide a powerful mechanism for regulating conflict in groups, even when there are more than two people and no central authority. This paper investigates the emergence and stability of behavioral norms in the context of a game played by people of limited rationality. The dynamics of this new norms game are analyzed with a computer simulation based upon the evolutionary principle that strategies shown to be relatively effective will be used more in the future than less effective strategies. The results show the conditions under which norms can evolve and prove stable. One interesting possibility is the employment of metanorms, *the willingness to punish someone who did not enforce a norm. Many historical examples of domestic and international norms are used to illustrate the wide variety of mechanisms that can support norms, including metanorms, dominance, internalization, deterrence, social proof, membership in groups, law, and reputation.*

An established norm can have tremendous power. This is illustrated by a historical instance of the norm of dueling. In 1804 Aaron Burr challenged Alexander Hamilton to a duel. Hamilton sat down the night before the duel was to take place and wrote down his thoughts. He gave five reasons against accepting the duel: his principles were against shedding blood in a private combat forbidden by law; he had a wife and children; he felt a sense of obligation toward his creditors; he bore no ill against Colonel Burr; and he would hazard much and could gain little. Moreover, he was reluctant to set a bad example by accepting a duel. Yet he did accept, because "the ability to be useful, whether in resisting mischief or effecting good, in those crises of our public affairs which seem likely to happen, would probably be inseparable from a conformity with public prejudice in this particular" (Truman, 1884, pp. 345–48). In other words, the prospect of sanctions imposed by the general public in support of dueling caused Hamilton to risk, and ultimately to lose, his life—a powerful norm indeed, and yet one that has all but disappeared today after centuries of power over life and death.

Today, norms still govern much of our political and social lives. In politics, civil rights and civil liberties are as much protected by informal norms for what is acceptable as they are by the powers of the formal legal system. Leadership is itself subject to the power of norms, as Nixon learned when he violated political norms in trying to cover up Watergate. The operation of Congress is shaped by many norms, including those governing reciprocity (Matthews, 1960) and apprenticeship (Krehbiel, 1985). Across many nations, tolerance of opposition is a fragile norm that has great impact on whether a democracy can survive in a given country (Almond and Verba, 1963; Dahl, 1966). In international political economy, norms are essential for the understanding of the operations of many

functional domains such as banking, oil, and foreign aid (Axelrod and Keohane, 1985; Keohane, 1984; Krasner, 1983). Even in the domain of power politics, norms have virtually wiped out colonialism, inhibited the use of chemical warfare, and retarded the spread of nuclear weapons.

Not only are norms important for many central issues in political science, but they are vital to the other social sciences as well. Sociology seeks to understand how different societies work, and clearly norms are important in these processes (e.g., Opp, 1979, 1983). Anthropology frequently deals with the unique features of various peoples by describing in great detail their practices and values, as in the case of feuding (e.g., Black-Michaud, 1975). Psychologists are concerned with how people influence each other and the manner in which an individual becomes socialized into a community (e.g., Darley and Batson, 1973; Sherif, 1936). Economists are becoming interested in the origin and operation of norms as they have come to realize that markets involve a great deal of behavior based on standards that no one individual can determine alone (e.g., Furubotn and Pejovich, 1974; Schotter, 1981).

Large numbers of individuals and even nations often display a great degree of coordinated behavior that serves to regulate conflict. When this coordinated behavior takes place without the intervention of a central authority to police the behavior, we tend to attribute the coordinated behavior and the resulting regulation of conflict to the existence of norms. To make this appeal to norms a useful explanation, we need a good theory of norms. Such a theory should help explain three things: how norms arise, how norms are maintained, and how one norm displaces another.

One of the most important features of norms is that the standing of a norm can change in a surprisingly short time. For example, after many centuries of colonialism, the intolerance of colonial dependence took hold in the relatively short period of just two decades after World War II. Before and after such a transition, the state of affairs seems very stable and perhaps even permanent. For this reason, awareness of a given norm is most intense precisely when it is being challenged. Examples of norms being challenged today include the right to smoke in public without asking permission, the use of gender-laden language, and the prohibition against the use of chemical warfare. Some of these challenges will succeed in establishing new norms, and some will fail altogether. Thus, what is needed is a theory that accounts not only for the norms existing at any point in time, but also for how norms change over time. To clarify these processes, one must first be clear about exactly what is being discussed.

In this next section the evolutionary approach to be used in this paper is explained. Following this, the results of computer simulations of the evolution of norms are presented. The computer simulations are then extended to include a specific mechanism for the enforcement of norms, called *metanorms*. After these formal models are investigated, a wide variety of processes that might help to sustain norms are discussed, along with suggestions about how they too can be modeled. The question of the origin and content of norms is considered, and finally, a summary and conclusion presents the findings of this paper in the broad context of social and political change.

The Evolutionary Approach

Norms have been defined in various ways in the different literatures and even within the same literature. The three most common types of definitions are based upon expectations, values, and behavior.

That these different definitions are used for the same concept reflects how expectations, values, and behavior are often closely linked. Definitions based upon expectations or values are favored by those who study norms as they exist in a given social setting. Such definitions are convenient because interviews can elicit the beliefs and values of the participants, whereas systematically observing their actual behavior is more difficult. Because for many purposes the most important thing is actual behavior, a behavioral definition will be used in this study.

DEFINITION. *A* norm *exists in a given social setting to the extent that individuals usually act in a certain way and are often punished when seen not to be acting in this way.*

This definition makes the existence of a norm a matter of degree, rather than an all or nothing proposition, which allows one to speak of the growth or decay of a norm. According to this definition, the extent to which a given type of action is a norm depends on just how often the action is taken and just how often someone is punished for not taking it.

To investigate the growth and decay of norms, I have formulated a norms game in which players can choose to defect and to punish those they have seen defecting. The goal of the investigation is to see when cooperation based upon emerging norms will develop. Ultimately, the purpose is to learn what conditions favor the development of norms so that cooperation can be promoted where it might not otherwise exist or be secure.

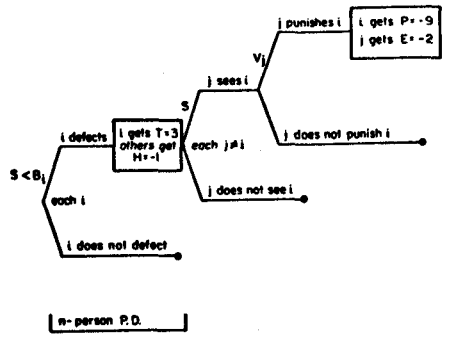
To see what rational actors would do in a particular setting, a game theory approach can be used. Game theory assumes the players are fully rational and choose the strategy that gives the highest expected utility over time, given their expectations about what the other players will do. Recent work by economists has shown great sophistication in dealing with problems of defining credible threats and of showing the consequences of requiring actors' expectations about each other to be consistent with the experience that will be generated by the resulting actions (Abreu, Pearce, and Stacchetti, 1985; Friedman, 1971; Kreps and Wilson, 1982; Selten, 1975).

While deductions about what fully rational actors will do are valuable for their own sake, empirical examples of changing norms suggest that real people are more likely to use trial and error behavior than detailed calculations based on accurate beliefs about the future. Therefore, I have chosen not to study the dynamics of norms using an approach that depends on the assumption of rationality.

Instead, I use an evolutionary approach. This approach is based on the principle that what works well for a player is more likely to be used again while what turns out poorly is more likely to be discarded (Axelrod, 1984). As in game theory, the players use their strategies with each other to achieve a payoff based upon their own choice and the choices of others. In an evolutionary approach, however, there is no need to assume a rational calculation to identify the best strategy. Instead, the analysis of what is chosen at any specific time is based upon an operationalization of the idea that effective strategies are more likely to be retained than ineffective strategies. Moreover, the evolutionary approach allows the introduction of new strategies as occasional random mutations of old strategies.

The evolutionary principle itself can be thought of as the consequence of any one of three different mechanisms. It could be that the more effective individuals are more likely to survive and reproduce. This is true in biological systems and in some economic and political systems. A second interpretation is that the players learn by trial and error, keeping effective strategies and altering ones that turn out

Figure 1. Norms Game



Key:

i, j	individuals
S	probability of a defection being seen by any given individual
B_i	boldness of i
V_j	vengefulness of j
T	player's temptation to defect
H	hurt suffered by others
P	cost of being punished
E	enforcement cost

poorly. A third interpretation, and the one most congenial to the study of norms, is that the players observe each other, and those with poor performance tend to imitate the strategies of those they see doing better. In any case, there is no need to assume that the individual is rational and understands the full strategic implications of the situation.

The evolutionary approach is inherently probabilistic and involves nonlinear effects. For these reasons, it is often impossible to use deductive mathematics to determine the consequences of a given model. Fortunately, computer simulation techniques (e.g., Cyert and March, 1963) provide a rigorous alternative to deductive mathematics. Moreover, simulation can reveal the dynamics of a process, as well as the equilibrium points. By simulating the choices of each member of a population of players and by seeing how the players' strategies change over time, the unfolding of a given evolutionary process can be analyzed to determine its overall implications.

The Norms Game

The norms game is described in Figure 1. It begins when an individual (i) has an opportunity to defect, say by cheating on an exam. This opportunity is accompanied by a known chance of being observed. The chance of being observed, or *seen*, is called S. If S is .5, each of the other players has an even chance of observing a defection if it takes place. If player i does defect, he or she gets a payoff of T (the *temptation* for defecting) equal to 3, and each of the others are *hurt* (H) slightly, getting a payoff of H equal to -1. If the player does not defect, no one gets anything.

So far the game is similar to an n-person Prisoner's Dilemma (see, e.g., G. Hardin, 1968; R. Hardin, 1982; Schelling, 1978). The new feature comes in the next step. If player i does defect, some of the other players may see the defection, and those who do may choose to punish the defector. If the defector is *punished* (P) the payoff is a very painful $P = -9$, but because the act of punishment is typically somewhat costly, the punisher has to pay an *enforcement cost* (E) equal to -2.

The strategy of a player thus has two dimensions. The first dimension of player i's strategy is *boldness* (B_i), which determines when the player will defect. The player will defect whenever the chance of being seen by someone is less than the player's boldness, which is to say, whenever $S < B_i$. The second dimension of a player's strategy is *vengefulness* (V_i), which is the probability that the player will punish someone who is defecting. The greater the player's vengefulness, the more likely he or she will be to punish someone who is spotted defecting.

Table 1. Example of Payoffs in the Norms Game Attained by a Player With Boldness Equal to 2/7 and Vengefulness Equal to 4/7

Event	Payoff per Event	Number of Events	Payoff
Defection	$T = 3$	1	3
Punishment	$P = -9$	1	-9
Hurt by others	$H = -1$	36	-36
Enforcement cost	$E = -2$	9	-18
Score			-60

Simulation of the Norms Game

The simulation of the norms game determines how the players' strategies evolve over time. The two dimensions of a strategy, boldness and vengefulness, are each allowed to take one of eight levels, from 0/7 to 7/7. Because the representation of eight levels requires three binary bits, the representation of a player's strategy requires a total of six bits, three for boldness and three for vengefulness.

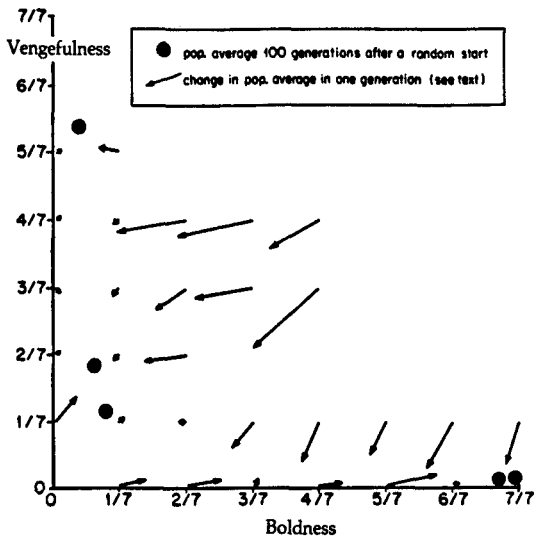
The simulation itself proceeds in five steps, as follows:

(1) The strategies for the initial population of 20 players are chosen at random from the set of all possible strategies.

(2) The score of each player is determined from the player's own choices and the choices of the other players. Each individual gets four opportunities to defect. For each of these opportunities, the chance of being seen, S, is drawn from a uniform distribution between 0 and 1. To see how the scores are attained, let us focus on an arbitrary player in the initial population of one of the runs, who will be called Lee. Lee has a boldness level of 2/7 and vengefulness level of 4/7. The total payoff Lee achieved was the result of four different kinds of events, as shown in Table 1. Lee defected only once because only one of the four opportunities had a chance of being seen that was less than Lee's boldness of 2/7. This defection gave a temptation payoff of $T = 3$ points. Unfortunately for Lee, one of the other players observed the defection and chose to punish it, leading to a loss for Lee of $P = -9$ points. In addition the other players defected a total of 36 times, each hurting Lee $H = -1$ point. Finally, Lee observed who was responsible for about half of these defections and chose to punish each of them with a probability determined by his vengefulness of 4/7. This lead to a punishment of 9 of the defections at an enforcement cost of $E = -2$ each, for a further loss of 18 points. The net result of these four types of events was a total score of -60 for Lee.

(3) When the scores of all the players are determined, individuals whose strategies were relatively successful are selected to have more offspring.[1] The method is to give an average individual one offspring and to give two offspring to an individual who is one standard deviation more effective than the average. An individual who is one standard deviation below the population average will not have his or her strategy reproduced at all. For convenience, the number of offspring is adjusted to maintain a constant population of 20. A final step is the introduction of some mutation so that new strategies can arise and be tested. This is done by allowing a 1% chance that each bit of an individual's new strategy will be altered. This mutation rate gives a little more than one mutation per generation in the entire population.

Figure 2. Norms Game Dynamics



(4) Steps 2 and 3 are repeated for 100 generations to determine how the population evolves.

(5) Steps 1 to 4 are repeated to give five complete runs of the simulation.

The results of the five runs are shown in Figure 2. The five circles indicate the average boldness and vengefulness of a population after 100 generations. Three completely different outcomes appear possible. In one of the runs, there was a moderate level of vengefulness and almost no boldness, indicating the partial establishment of a norm against defection. On two other runs there was little boldness and little vengefulness, and on the remaining two runs, there was a great deal of boldness and almost no vengefulness—the very opposite of a norm against defection. What could be happening?

The way the strategies actually evolve over time is revealed by the change that takes place in a single generation in a population's average boldness and vengefulness. To calculate this, the data are used from all 100 generations of all five runs, giving 500 populations. The populations with similar average boldness and vengefulness are then grouped together, and their average boldness and vengefulness one generation later is measured. The results are indicated by the arrows in Figure 2.

Now the various outcomes begin to fit into a common pattern. All five of the runs begin near the middle of the field, with average boldness and vengefulness levels near one-half. The first thing to happen is a dramatic fall in the boldness level. The reason for the decline is that when there is enough vengefulness in the population, it is very costly to be bold. Once the boldness level falls, the main trend is a lowering of vengefulness. The reason for this is that to be vengeful and punish an observed defection requires paying an enforcement cost without any direct return to the individual. Finally, once the vengefulness level has fallen nearly to zero, the players can be bold *with impunity*. This results in an increase in boldness, destroying whatever restraint was established in the first stage of the process—a sad but stable state in this norms game.

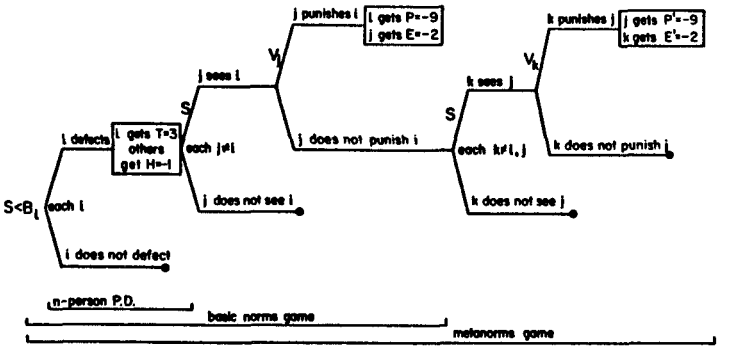
This result raises the question of just what it takes to get a norm established. Because the problem is that no one has any incentive to punish a defection, the next section explores one of the mechanisms that provides an incentive to be vengeful.

Metanorms

A little-lamented norm of once great strength was the practice of lynching to enforce white rule in the South. A particularly illuminating episode took place in Texas in 1930 after a black man was arrested for attacking a white woman. The mob was impatient, so they burned down the courthouse to kill the prisoner within. A witness said,

> I heard a man right behind me remark of the fire, "Now ain't that a shame?" No sooner had the

Figure 3. Metanorms Game



Key:

i,j,k individuals
S probability of a defection's being seen by any given individual
B_i boldness of i
V_j vengefulness of j
T temptation to defect
H hurt suffered by others
P cost of being punished
E punisher's enforcement cost
P' cost of being punished for not punishing a defection
E' cost of punishing someone for not punishing a defection

> words left his mouth than someone knocked him down with a pop bottle. He was hit in the mouth and had several teeth broken. (Cantril, 1941, p. 101)

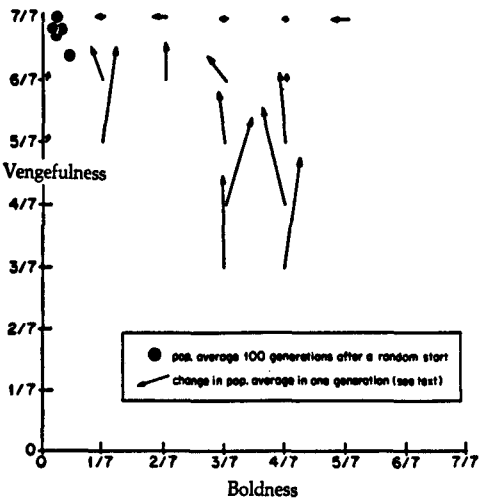
This is one way to enforce a norm: punish those who do not support it. In other words, be vengeful, not only against the violators of the norm, but also against anyone who refuses to punish the defectors. This amounts to establishing a norm that one must punish those who do not punish a defection. This is what I will call a *metanorm.*

Metanorms are widely used in the systems of denunciation in communist societies. When the authorities accuse someone of doing something wrong, others are called upon to denounce the accused. Not to join in this form of punishment is itself taken as a defection against the group (Bronfenbrenner, 1970; Meyers and Bradbury, 1968).

As another example, when the Soviet Union supported the suppressionn of the Solidarity movement in Poland, the United States asked its allies to stop supplying components to the Soviet Union for its new gas pipeline. The allies, not wanting to pay the enforcement cost of this punishment, refused. The United States government then undertook the metapunishment of imposing sanctions on foreign companies that defied the sales ban (*New York Times,* January 5 and June 19, 1982).

The formulation of a metanorms game can help in the exploration of the effectiveness of this mechanism. Figure 3 shows how the metanorms game is based upon an extension of the norms game. If

Figure 4. Metanorms Game Dynamics



someone defects, and Lee sees but does not punish that defection, then the other players have a chance to see and punish Lee. The model makes the critical assumption that a player's vengefulness against nonpunishment is the same as the player's vengefulness against an original defection.[2] The validity of this assumption will be addressed later, but first let us see what affect it has on the evolution of the process.

A set of five runs was conducted with the metanorms game, each done as before with a population of 20 players and a duration of 100 generations. The results are shown in Figure 4. They are unambiguous. In all five runs a norm against defection was established. The dynamics are clear. The amount of vengefulness quickly increased to very high levels, and this in turn drove down boldness. The logic is also clear. At first there was a moderate amount of vengefulness in the population. This meant that a player had a strong incentive to be vengeful, namely, to escape punishment for not punishing an observed defection. Moreover, when each of the players is vengeful out of self-protection, it does not pay for anyone to be bold. Thus the entire system is self-policing, and the norm becomes well established.

This result is dependent, however, on the population's starting with a sufficiently high level of vengefulness. Otherwise the norm still collapses. Thus, while the norms game collapses no matter what the initial conditions are, the metanorms game can prevent defections if the initial conditions are favorable enough.

Mechanisms to Support Norms

The simulations of the norms game and the metanorms game have allowed the exploration of some of the important processes in the dynamics of norms. The simulation of the norms game shows that relying on individuals to punish defections may not be enough to maintain a norm. Therefore, the question to be considered now is, What mechanisms can serve to support a norm that is only partially established? The evolutionary approach helps to develop a list of such processes, and in some cases, suggests specific methods for modeling the process by which a norm can be supported.

Metanorms

As the computer simulations show, the existence of a metanorm can be an effective way to get a norm started and to protect it once it is established. By linking vengefulness against nonpunishers with vengefulness against defectors, the metanorm provides a mechanism by which the norm against defection becomes self-policing. The trick, of course, is to link the two kinds of vengefulness. Without this link, the system could unravel. An individual might reduce the metavengeance level while still being vengeful and then later stop being vengeful when others stopped being metavengeful.

The examples cited earlier suggest that

people may well punish those who do not help to enforce a valued norm. The model suggests norms can be supported if people tend to have correlated degrees of vengefulness or anger against someone who violates a particular norm and someone who tolerates such a violation. What the evolutionary approach has done is raise the possibility that metanorms are a mechanism that can help support norms, thus suggesting the interesting empirical question of whether the two types of vengefulness are indeed correlated. My guess is that there is such a correlation. The types of defection we are most angry about are likely to be the ones whose toleration also makes us angry. As of now, however, the possibility of metanorms remains speculative.

Dominance

Another mechanism for supporting a norm is the dominance of one group over another. For example, it is no coincidence that in the South, whites lynched blacks, but blacks did not lynch whites. The whites had two basic advantages: greater economic and political power, and greater numbers.

Simulation of the effects of power and numbers can be readily done with slight extensions of the basic model to allow for the existence of two different groups. The competition between two groups can be modeled by assuming that the defections of a player only hurt the members of the other group and are therefore only punished by members of the other group. Similarly, in the metanorms version of the model, punishments for not punishing a defector would only occur within a group, as illustrated by the pop bottle used by one white against another in the lynching example discussed above. Moreover, in determining strategies for the next generation, the strategies of two groups would be allowed to adapt separately so that whites learn from whites and blacks learn from blacks.

The two advantages of the whites are modeled separately. Their greater economic or political power is reflected in their lessened cost of being punished by a black. This was done by letting $P = -3$ for whites while retaining $P = -9$ for blacks. The greater numbers of whites are reflected directly in the relative size of the two populations, giving the whites a greater chance to observe and punish a black defection than vice versa. This was done by letting the population be 20 whites and 10 blacks.

Analysis of runs based upon these conditions shows that resistance to punishment and increased size can help a group, but only if there are metanorms. Without metanorms, even members of the stronger group tend to be free riders, with no private incentive to bear enforcement costs. This in turn leads to low vengefulness and high boldness in both groups. When metanorms are added, it becomes relatively easier for the strong group to keep the weak group from being bold, while it is not so easy for the weak group to keep the strong one from defecting.

Another form of potential strength is illustrated by the case of a major power interacting with many smaller nations. For example, the U.S. may not only be in a favorable position on a given bilateral interaction but also may have many more bilateral interactions than others. Thus, its behavior has a greater impact on the development of norms than would the behavior of a minor power. When Libya wanted to modify the international norm of the twelve-mile limit of territorial waters to include the entire Gulf of Sidra, the United States fleet deliberately sailed into the Gulf and shot down two Libyan planes sent up to try to change the norm. Clearly the U.S. was not only stronger but had incentives to enforce the old norm based upon its naval interests in other parts of the world.

While the process of frequent interactions by a single strong player has not

yet been simulated, it is plausible that such a process would help to establish a norm against defection because the central player would have a greater unilateral incentive to be vengeful against defections.

Norms can also be promoted by the interests of a few major actors, such as the U.S. and the Soviet Union's both working to retard the proliferation of nuclear weapons. Their actions need not be coordinated in detail as long as together they are important enough to others to enforce a norm of the major actors' choice. The logic is somewhat analogous to Olson's "privileged group" in a collective action problem (Olson, 1965, pp. 48–50).

Internalization

Norms frequently become internalized (Scott, 1971). This means that violating an established norm is psychologically painful even if the direct material benefits are positive. This is frequently observed in laboratory experiments where subjects are more equitable than they have to be and explain their behavior by saying things like "you have to live with yourself." In terms of the norms game, this type of internalization means that the temptation to defect, T, is negative rather than positive. If everyone internalizes a given norm this strongly, there is no incentive to defect and the norm remains stable. Obviously families and societies work very hard to internalize a wide variety of norms, especially in the impressionable young. They do so with greater or lesser success depending on many factors, including the degree to which the individual identifies with the group and the degree to which the norm and its sponsors are seen as legitimate.[3]

Clearly, it is rare for everyone in a group to have a norm so strongly internalized that for each the temptation to defect is actually negative. An interesting question for future modeling is, How many people have to internalize a norm in order for it to remain stable?

The logic of the norms game suggests that lowering the temptation to defect might not be enough. After all, even if most people did not defect, if no one had an incentive to punish the remaining defectors, the norm could still collapse. This point suggests that we look for internalization, not only in the reduced incentive to defect, but also in an increased incentive to punish someone else who does defect.

An increased incentive to punish, through internalization or by some other means, would lead some people to feel a gain from punishing a defector. For them, the payoff from enforcement, E, would actually be positive. Such people are often known as self-righteous busy bodies and often are not very well liked by those who enjoy a defection now and then. Given enough people who enjoy enforcing the norm, the question of its maintenance then becomes whether the chance is high or low that the defection will be seen.

Deterrence

In the norms game and the metanorms game the players do not look ahead. Instead they try a particular strategy, see how it does, compare their payoff with the payoff of others, and switch strategies if they are doing relatively poorly. While trial and error is a sensible way of modeling players of very limited rationality, it does not capture the idea that players may have a great enough understanding of the situation to do some forward-looking calculations as well as backward-looking comparisons with others. In particular, a person may realize that even if punishing a defection is costly now, it might have long-term gains by discouraging other defections later.

A good example is the strong U.S. response to New Zealand's refusal in February 1985 to allow a U.S. destroyer into

Aukland harbor without assurances that it did not carry nuclear weapons. The U.S. government presumably did not care very much about nuclear access to New Zealand ports, but it did care a great deal about deterring the spread of a new norm of "nuclear allergy" among its many allies in other parts of the world (Arkin and Fieldhouse, 1985).

Social Proof

An important principle from social psychology is "social proof," which applies especially to what people decide is correct behavior. As Cialdini (1984, p. 117) explains,

> we view a behavior as more correct in a given situation to the degree that we see others performing it. Whether the question is what to do with an empty popcorn box in a movie theater, how fast to drive on a certain stretch of highway, or how to eat chicken at a dinner party, the actions of those around us will be important in defining the answer.

The actions of those around us serve several functions. First, they provide information about the boldness levels of others, and indirectly about the vengefulness of the population. Hence, we can infer something about whether it pays for us to be bold or not. Second, the actions of others might contain clues about what is the best course of action even if there is no vengefulness. For example, people may be driving slowly on a certain stretch of highway, not because there is a speed trap there, but because the road is poorly paved just ahead. Either way, the actions of others can provide information about how the population has been adapting to a particular environment. If we are new to that environment, this is valuable information about what our own behavior should be (Asch, 1951, 1956; Sherif, 1936). The actions of others provide information about what is proper for us, even if we do not know the reasons. Finally, in many cases, by conforming to the actions of those around us, we fulfill a psychological need to be part of a group.

Our propensity to act on the principle of social proof is a major mechanism in the support of norms. The current model of norms already has a form of this mechanism built in: when a relatively unsuccessful individual seeks a new strategy, that strategy is selected from those being used by the rest of the population. This is a form of social proof, refined by giving weight to the more successful strategies being employed in the population.

In cases where other people differ in important ways, the principle of social proof tends to apply to those who are most like us. This too is easy to build into simulations with more than one group. In the simulation of blacks and whites, the blacks look only to other blacks when selecting a new strategy, and the whites look only to other whites. This makes good sense because a strategy that is very successful for a white might be disastrous if employed by a black.

Membership

Another mechanism for the support of norms is voluntary membership in a group working together for a common end.[4] Contracts, treaties, alliances, and memberships in social groups all carry with them some power to impose obligations upon individuals. The power of the membership works in three ways. First, it directly affects the individual's utility function, making a defection less attractive because to defect against a voluntarily accepted commitment would tend to lower one's self-esteem. Second, group membership allows like-minded people to interact with each other, and this self-selection tends to make it much easier for the members to enforce the norm implicit in the agreement to form or join a group. Finally, the very agreement to form a group helps define what is expected of the participants, thereby clarifying when a

defection occurs and when a punishment is called for.

One might suppose it would be easy for a bold individual to join and then exploit a group that had gathered together in the expectation of mutual compliance. Actually, this does not usually happen, in part because the factors just outlined tend to isolate a defector and make it relatively easy for the others to be vengeful—especially with the help of metanorms. Another factor is that, according to recent experimental evidence, cooperators are more likely to stay in a group than are defectors (Orbell, Schwarz-Shea, and Simmons, 1984). This happens because cooperators have a stronger ethical or group-regarding impulse than defectors, a factor that led them to cooperate in the first place.

The metanorms game can be expanded to include the choice of whether to join a group or not.[5] In general, the value to a person of joining a group would depend on how many others joined. Each player would make this choice at the start of the game. Then the interactions concerning defections and punishments would occur as before, with the interactions limited to those who had actually joined. As an example, an alliance for collective security would include a group of nations that had joined for this common purpose. Once a nation had joined, a defection would consist of not supporting the alliance in some collective security task. A defection would hurt the other members of the alliance, and some of them might choose to punish the defector; they might also choose to punish someone who did not punish the defector. Typically, the larger the number of nations joining the group, the greater the benefits of cooperation would be for its members.

In the political sphere, voluntary membership taking the form of a social contract has been a powerful image for the support of democratic forms of governance. In effect, a mythical agreement is used to give legitimacy to a very real set of laws and institutions.

Law

Norms often precede laws but are then supported, maintained, and extended by laws. For example, social norms about smoking in public are now changing. As more and more people turn vengeful against someone who lights up in a confined space, fewer and fewer smokers are so bold as to do so without asking permission. As this norm becomes firmer, there is growing support to formalize it through the promulgation of laws defining where smoking is and is not permitted.[6]

A law supports a norm in several ways. The most obvious is that it supplements private enforcement mechanisms with the strength of the state. Because enforcement can be expensive for the individual, this can be a tremendous asset. In effect, under the law the collective goods problem of enforcement is avoided because selective incentives are given to specialized individuals (inspectors, police, judges, etc.) to find and punish violations.

The law also has a substantial power of its own, quite apart from whether it is or can be enforced. Many people are likely to take seriously the idea that a specific act is mandated by the law, whether it is a requirement to use seat belts or an income tax on capital gains. However, we all know this respect for the law has its limits, and we suspect that many people do not pay all the tax they should. Even when enforcement is possible and is attempted, the strength of the law is limited. In most cases, the law can only work as a supplement (and not a replacement) for informal enforcement of the norm. The failure of Prohibition is a classic example of an attempt to enforce a norm without sufficient social support.

In addition to enforcement and respect, a third advantage of the law is clarity. The law tends to define obligations much

more clearly than does an informal norm. A social norm might say that a landlord should provide safe housing for tenants, but a housing code is more likely to define safety in terms of fire escapes. Over the domain covered by the law, the norm might become quite clear. However, this clarity is gained at the expense of suggesting that conformity with the law is the limit of one's social obligations.

Modeling the power and operation of the law is beyond the scope of this project. However, it should still be emphasized that often law is the formalization of what has already attained strength as a social or political norm. An important example is civil liberties, the very foundation of a democratic system. There are laws and constitutional provisions in support of civil liberties such as freedom of speech, but the legal system can only protect free speech if there is substantial support for it among a population willing to tolerate dissent and willing to protect those who exercise it.

In short, social norms and laws are often mutually supporting. This is true because social norms can become formalized into laws and because laws provide external validation of norms. They are also mutually supporting because they have complementary strengths and weaknesses. Social norms are often best at preventing numerous small defections where the cost of enforcement is low. Laws, on the other hand, often function best to prevent rare but large defections because substantial resources are available for enforcement.

Reputation

An important, and often dominant, reason to respect a norm is that violating it would provide a signal about the type of person you are. For example, if there is a norm dictating that people should dress formally for dinner, and you don't, then others might make some quite general inferences about you.

The importance of dressing formally when the occasion requires is not just that others will punish you for violating the norm (say, by giving you a disapproving look) but also that they will infer things about you and then act in ways you wish they wouldn't. This is an example of the signaling principle: a violation of a norm is not only a bit of behavior having a payoff for the defector and for others; it is also a signal that contains information about the future behavior of the defector in a wide variety of situations.[7]

There are several important implications of the signaling principle for the origin and durability of a norm. A norm is likely to originate in a type of behavior that signals things about individuals that will lead others to reward them. For example, if a certain accent signals good breeding, then others may give better treatment to those who speak that way. Once this happens, more people are likely to try to speak that way. Eventually, people might be punished (e.g., despised) for not having the right accent. Thus, what starts out as a signal about one person's background can become a norm for all.[8]

The signaling principle helps explain how an "is" becomes an "ought." As more and more people use the signal to gain information about others, more and more people will adopt the behavior that leads to being treated well. Gradually the signal will change from indicating a rare person to indicating a common person. On the other hand, the absence of the signal, which originally carried little information, will come to carry substantial information when the signal becomes common. When almost everyone behaves in conformity with a signal, those who don't stand out. These people can now be regarded as violators of a norm—and dealt with accordingly.

Note that there is an important distinction between a convention, which has no

direct payoffs one way or the other (such as wearing a tie for men), and a cooperative act, the violation of which leads to injury to others (e.g., queuing for service). A type of behavior with no direct payoffs can become a norm once it develops some signaling value, as is the case when fashion leaders adopt a new style (Veblen, 1899). Once this happens, a violator of this style will be looked down upon. Thus the style will become a norm; individuals will usually follow the style, and those who do not will likely be punished.

The Origin and Content of Norms

Eight mechanisms have now been identified that can serve to support a norm that is already at least partially established. What, however, are the characteristics of the behaviors that arise and then become more and more established as norms? Or to put it another way, just what is the content of behavior that might later turn into a norm?

The answer depends on what types of behavior can appear and spread in a population even when only a few people initially exhibit the behavior. This, in turn, depends on what kind of behavior is likely to be rewarded and punished for its own sake, independently of whether or not it is common behavior.

Two of the supporting mechanisms already considered can serve in this initial role: dominance and reputation. Dominance can work because if only a few very powerful actors want to promote a certain pattern of behavior, their punishments alone can often be sufficient to establish it, even if the others are not vengeful against defections. The implications for the substance of norms are obvious: it is easier to get a norm started if it serves the interests of the powerful few.

In fact, many norms obeyed and even enforced by almost everyone actually serve the powerful. This can happen in forms disguised as equalitarian or in forms that are blatantly hierarchical. An apparently equalitarian norm is that the rich and the poor are equally prohibited from sleeping under bridges at night. A blatantly hierarchical norm is that soldiers shall obey their officers. Both forms are "norms of partiality," to use the term of Ullman-Margalit (1977).

To say that the powerful can start a norm suggests a great deal about the potential substance of such norms. Once started, the strong support the norms because the norms support the strong.

Dominance is not the only mechanism capable of starting a norm. Reputation can do so as well. Consider, for example, the idea of keeping one's promise. In a hypothetical society in which few people kept their promises, you would be happy to deal with someone who did. You would find it in your narrow self-interest to continue dealing with such a person, and this in turn would be rewarding to the promise-keeper. Conversely, you would try to avoid deals with those you knew did not keep their promises. You would, in effect, be vengeful against defectors without having to pay an enforcement cost. Indeed, your enforcement would simply be the result of your acting in your own interests, based upon the reputations of others and your calculation about what was good for yourself.

International regimes depend on just such reputational mechanisms to get norms started (Keohane, 1984). In such cases, countries can be very deliberate about what promises they make and which ones they want to keep when the stakes are high (Axelrod, 1979). Reputational effects can also be based upon the limited rationality of trial and error learning. If a person associates another's response to a particular act (say a refusal to continue dealing as a reaction to the breaking of a promise), then the violator can learn not to break promises.

This learning approach suggests the

importance of being able to link the behavior with the response. Behaviors will be easier to establish as norms if the optimal response of others is prompt and rewarding. Failing a prompt response, learning can also take place if the delayed punishment is explicitly cited as a response to the earlier defection.

Summary and Conclusion

To study the development of norms, the strategic situation has been modeled as an *n*-person game. In the basic norms game, everyone has two types of choice: the choice to cooperate or defect, which affects everyone, and the choice of whether or not to punish a specific person seen defecting. A player's strategy is described in terms of how these choices will be made. A strategy consists of two parameters: boldness (the largest chance of being seen that will lead to a choice of defection) and vengefulness (the probability of punishing someone observed defecting). To the extent that players are vengeful, but not very bold, a norm can be said to have been established.

To study the dynamics of the process, an evolutionary approach was employed. In this approach, the initial strategies are chosen at random, and the population of players is given opportunities to defect and to punish the defections they observe. The evolutionary approach dictates that strategies proving relatively effective are more likely to be employed in the future while less effective strategies are dropped. Moreover, strategies undergo some random mutation so that new ones are always being introduced into the population.

The computer simulation of this process revealed an interesting dynamic in the norms game. At first, boldness levels fell dramatically due to the vengefulness in the population. Then, gradually, the amount of vengefulness also fell because there was no direct incentive to pay the enforcement cost of punishing a defection. Once vengeance became rare, the average level of boldness rose again, and the norm completely collapsed. Moreover, the collapse was a stable outcome.

This result led to a search for mechanisms that could sustain a partially established norm. One possibility is the metanorm: the treatment of nonpunishment as if it were another form of defection; that is, a player will be vengeful against someone who observed a defection but did not punish it. Simulation of the evolution of strategies in this metanorms game demonstrated that players had a strong incentive to increase their vengefulness lest they be punished by others, and this in turn led to a decline of boldness. Thus, metanorms can promote and sustain cooperation in a population.

Other mechanisms for the support of norms are also important. These include dominance, internalization, deterrence, social proof, membership, law, and reputation. In some cases, the resulting norms are hierarchical rather than equalitarian, and the cooperation exhibited is coerced rather than freely offered. A good example is the norm of black deference in the old South.

Dominance processes have been simulated by subdividing the population and letting one segment be relatively resistant to the effects of punishment by members of the other segment. Internalization can be investigated by studying the effects of making defection costly rather than rewarding for some of the defectors and by making punishment a pleasure rather than a cost for some of the observers of a defection. A more drastic change in the modeling procedures would be necessary to study some of the other mechanisms in question.

Norms are important in society and, not surprisingly, have been given a great deal of attention in the social sciences, including sociology, anthropology, political science, psychology, and economics.

While descriptions of actual norms abound, investigations of the reasons for people to obey or violate a given norm have been much less common. Even among the strategic approaches to norms, relatively little attention has been devoted to understanding the dynamics of norms: how they can get started, how a partial norm can be sustained and become well established, and how one norm can displace another. An evolutionary approach is helpful in studying these dynamics because it can help show how strategies change over time as a function of their relative success in an ever-changing environment of other players who are also changing their own strategies with experience.

A major goal of investigating how cooperative norms in societal settings have been established is a better understanding of how to promote cooperative norms in international settings. This is not as utopian as it might seem because international norms against slavery and colonialism are already strong while international norms are partly effective against racial discrimination, chemical warfare, and the proliferation of nuclear weapons. Because norms sometimes become established surprisingly quickly, there may be some useful cooperative norms that could be hurried along with relatively modest interventions.

Notes

I owe a great deal to Stephanie Forrest, my research assistant, and to those who helped me think about norms: Michael Cohen, Jeffrey Coleman, John Ferejohn, Morris Fiorina, Robert Gilpin, Donald Herzog, John Holland, Melanie Manion, Ann McGuire, Robert Keohane, Robert McCalla, Amy Saldinger, Lynn Sanders, Kim Scheppele, Andrew Sobel, Charles Stein, Laura Stoker, and David Yoon. I am also pleased to thank those who helped support various aspects of this work: the Harry Frank Guggenheim Foundation, the National Science Foundation, the Sloan Foundation, and the Michigan Memorial Phoenix Project.

1. The procedure used is inspired by the genetic algorithm of computer scientist John Holland (1975, 1980).

2. For convenience, it is also assumed that the chance of being seen not punishing is the same as the chance of the original defection being seen. The payoff for metapunishment is $P' = -9$, and the metaenforcement cost is $E' = -2$.

3. Marx goes as far as to say that social norms are merely reflections of the interests of the ruling class, and the other classes are socialized into accepting these norms under "false consciousness."

4. I thank David Yoon and Lynn Sanders for pointing this out to me.

5. I thank David Yoon for formulating this variant of the metanorms game and the application to alliances that follows.

6. The same process of formalizing norms applies to private laws and regulations, as in the case of a business that issues an internal rule about who is responsible for making coffee.

7. For the theory of signaling, see Spence (1974). For a theory of how customs can be sustained by reputations, see Akerlof (1980).

8. Signals can also help to differentiate groups and thereby maintain group boundaries and cohesiveness.

References

Abreu, Dilip, David Pearce, and Ennio Stacchetti. 1985. Optimal Cartel Equilibria with Imperfect Monitoring. Minneapolis: University of Minnesota Institute for Mathematics and its Applications.

Akerlof, George A. 1980. A Theory of Social Custom, of Which Unemployment May Be One Consequence. *Quarterly Journal of Economics*, 94:749–75.

Almond, Gabriel, and Sidney Verba. 1963. *The Civic Culture*. Princeton: Princeton University Press.

Arkin, William, and Richard W. Fieldhouse. 1985. Focus on the Nuclear Infrastructure. *Bulletin of the Atomic Scientists*, 41:11–15.

Asch, Solomon E. 1951. Effects of Group Pressure upon the Modification and Distortion of Judgment. In Harold Guetzkow, ed., *Groups, Leadership and Men*. Pittsburgh: Carnegie Press.

Asch, Solomon E. 1956. Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority. *Psychological Monographs*, 41:258–90.

Axelrod, Robert. 1979. The Rational Timing of Surprise. *World Politics*, 31:228–46.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert, and Robert O. Keohane. 1985. Achieving Cooperation Under Anarchy: Strategies and Institutions. *World Politics*, 38:226–54.

Black-Michaud, Jacob. 1975. *Cohesive Force: Feud in the Mediterranean and the Middle East.* Oxford: Basil Blackwell.

Bronfenbrenner, Urie. 1970. *Two Worlds of Childhood: U.S. and U.S.S.R.* New York: Russell Sage.

Cantril, Hadley. 1941. *The Psychology of Social Movements.* New York: Wiley.

Cialdini, Robert H. 1984. *Influence—How and Why People Agree to Things.* New York: Morrow.

Cyert, Richard M., and James G. March. 1963. *A Behavioral Theory of the Firm.* Englewood Cliffs, NJ: Prentice-Hall.

Dahl, Robert A., ed. 1966. *Political Oppositions in Western Democracies.* New Haven: Yale University Press.

Darley, John M., and C. Daniel Batson. 1973. "From Jerusalem to Jerico": A Study of Situational and Dispositional Variables in Helping Behavior. *Journal of Personality and Social Psychology,* 27:100–108.

Friedman, James W. 1971. A Non-cooperative Equilibrium for Supergames. *Review of Economic Studies,* 38:1–12.

Furubotn, Eirik G., and Svetozar Pejovich, eds. 1974. *The Economics of Property Rights.* Cambridge, MA: Ballinger.

Hardin, Garrett. 1968. The Tragedy of the Commons. *Science,* 162:1243–48.

Hardin, Russell. 1982. *Collective Action.* Baltimore: Johns Hopkins University Press.

Holland, John H. 1975. *Adaptation in Natural and Artificial Systems.* Ann Arbor: University of Michigan Press.

Holland, John H. 1980. Adaptive Algorithms for Discovering and Using General Patterns in Growing Knowledge Bases. *International Journal of Policy Analysis and Information Systems,* 4: 245–68.

Keohane, Robert O. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy.* Princeton: Princeton University Press.

Krasner, Stephen D., ed. 1983. *International Regimes.* Ithaca: Cornell University Press.

Krehbiel, Keith. 1985. Unanimous Consent Agreements: Going Along in the Senate. Working paper no. 568. California Institute of Technology, Social Science Department.

Kreps, David M., and Robert Wilson. 1982. Sequential Equilibria. *Econometrica,* 50:863–94.

Matthews, Donald R. 1960. *U.S. Senators and Their World.* Chapel Hill: University of North Carolina Press.

Meyers, Samuel M., and William C. Bradbury. 1968. The Political Behavior of Korean and Chinese Prisoners of War in the Korean Conflict: A Historical Analysis. In Samuel M. Meyers and Albert D. Briderman, eds., *Mass Behavior in Battle and Captivity, The Communist Soldier in the Korean War.* Chicago: University of Chicago Press.

Olson, Mancur. 1965. *The Logic of Collective Action.* Cambridge, MA: Harvard University Press.

Opp, Karl-Dieter. 1979. Emergence and Effects of Social Norms—Confrontation of Some Hypotheses of Sociology and Economics. *Kylos,* 32: 775–801.

Opp, Karl-Dieter. 1983. Evolutionary Emergence of Norms. *British Journal of Social Psychology,* 21: 139–49.

Orbell, John M., Peregrine Schwartz-Shea, and Randall T. Simmons. 1984. Do Cooperators Exit More Readily than Defectors? *American Political Science Review,* 78:163–78.

Schelling, Thomas. 1978. *Micromotives and Macrobehavior.* New York: W. W. Norton.

Schotter, Andrew. 1981. *Economic Theory of Social Institutions.* Cambridge: Cambridge University Press.

Scott, John F. 1971. *Internalization of Norms.* Englewood Cliffs, NJ: Prentice-Hall.

Selten, R. 1975. Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory,* 4:25–55.

Sherif, Muzafer. 1936. *The Psychology of Social Norms.* New York: Harper and Brothers.

Spence, A. Michael. 1974. *Market Signalling.* Cambridge, MA: Harvard University Press.

Truman, Ben C. 1884. *Field of Honor: A Complete and Comprehensive History of Dueling in All Countries.* New York: Fords, Howard and Hilbert.

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

Veblen, Thorstein. 1899. *The Theory of the Leisure Class.* New York: Macmillan.

Robert Axelrod is Professor of Political Science, University of Michigan, Ann Arbor, MI 48109.