J. theor. Biol. (2003) **220,** 345–357
doi:10.1006/jtbi.2003.3166, available online at http://www.idealibrary.com on IDEAL®

Reputation and the Evolution of Conflict

RICHARD MCELREATH*†‡

†*Department of Anthropology, University of California, Davis, One Shields Avenue, Davis, CA* 95616-8522, *U.S.A. and* ‡*Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee* 94, 14195 *Berlin, Germany*

(*Received on* 9 *November* 2001, *Accepted in revised form on* 16 *September* 2002)

The outcomes of conflicts in many human societies generate reputation effects that influence the nature of later conflicts. Those willing to escalate over even trivial offenses are considered honorable whereas those who do not are considered dishonorable (Nisbett & Cohen, 1996). Here I extend Maynard Smith's hawk–dove model of animal conflict to explore the logic of a strategy which uses reputation about its opponents to regulate its behavior. I show that a reputation-based strategy does well when (1) the value of the resource is large relative to the cost of losing a fight, (2) communities are stable, and (3) reputations are well known but subject to some amount of error. Reputation-based strategies may thus result in greater willingness to fight, but less fighting at equilibrium, depending upon the nature of the contests and the local socioecology. Additionally, this strategy is robust in the presence of poor knowledge about reputation.

© 2003 Elsevier Science Ltd. All rights reserved.

Introduction

People in many societies participate in what Nisbett & Cohen (1996) have called "cultures of honor," engaging in costly fights over seemingly trivial matters in order to preserve their public standing. Examples of cultures of honor are and have been present in the American south, the American west, the circum-Mediterranean, and parts of Africa. Nisbett and Cohen argue that cultures of honor are likely to develop and be stable in environments where there is little law enforcement (such as frontier communities) and wealth is easy to lose (such as herding economies, where wealth is mobile and therefore easy to steal). As a result, individuals are forced to establish a reputation as a fighter in order to protect themselves from violence and theft. Essentially, a reputation as an individual who is willing to engage in costly fights serves as a deterrent to those who might otherwise attempt to seize property from or do violence to him and his family. This then allows the honorable individual to avoid future costly contests and instead reach peaceful conclusions to conflicts with other honorable men, since each believes that the other would readily fight to defend his claim.

The logic of the above argument has not yet been formalized. Despite interest in reputation and public image for regulating cooperation (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998a,b; Panchanathan & Boyd,

*Corresponding author. Department of Anthropology, University of California, Davis, One Shields Avenue, Davis, CA 95616-8522, U.S.A. Fax: +1-530-752-8885.
E-mail address: mcelreath@davis.edu (R. McElreath).

0022-5193/03/$35.00
© 2003 Elsevier Science Ltd. All rights reserved.

under review; Sugden, 1986), there has been little work on formal models of reputation in conflict. However, Johnstone (2001), independent of the model in this paper (McElreath, 2001), recently derived a model of "eavesdropping" in the hawk–dove game. Johnstone's model tackles a related problem, assuming that individuals attend to outcomes of fights (who won and who lost) in order to decide who is and is not willing to fight over a resource. However, several of the structural assumptions, like the absence of memory and the labeling of interactions between doves as conflicts, make it difficult to see the eavesdropper strategy as a good caricature of the economics of human reputation.

Here I present a game theoretic model of conflict regulated by a reputation for willingness to fight. An "honorable" individual is one in "good" standing since he is always willing to defend himself (play hawk) against aggression, even when such defense is on average costly to the individuals involved. What differentiates this strategy from a simple retaliatory strategy is that individuals also keep track of the reputations of others, and then use those reputations to decide how to respond when contests arise. With other individuals in good standing, they seek nonviolent resolutions (play dove). With those who have shown themselves to be without "honor" (by playing dove when an opponent plays hawk), they take advantage of that knowledge by attempting to take advantage of the person (by playing hawk). This model generates very different equilibria than Johnstone's model of eavesdropping. Whereas the eavesdropping model results in more conflict at equilibrium, this model may result in much less conflict, once individuals who use reputation to regulate contests become common. I show that the honorable strategy, "tough," comes to dominate the population when (1) the value of the resource is large relative to the cost of losing a fight, (2) communities are stable, and (3) reputations are well known or easy to assess but subject to small amounts of error. I discuss the evolutionary stability of the parameters and how the model reflects on societies in which cultures of honor are present or absent. Finally, I suggest some future directions for modeling in this area.

A Simple Model of Conflict and Reputation

The game proceeds in discrete time periods or generations. Each generation, pairs of individuals, sampled at random from an infinite population, interact. Interactions involve a contest over a divisible resource with value v. If the individuals fight over the resource, each has an equal chance of losing, and the loser pays a cost $-c$, while the victor receives v. Furthermore, $0 < v < c$ so that escalated contests are on average costly. If the individuals agree to settle the dispute peacefully, each receives half the value of the resource, $v/2$. If one individual refuses to fight, her payoff is unchanged while her opponent receives v, the value of the resource. Table 1 summarizes these payoffs, which constitute the familiar hawk–dove (or Chicken) game (Maynard Smith & Price, 1973).

Each round after the first, there is a probability w that an individual goes on to another interaction with another randomly selected individual. Also, individuals never interact with the same person twice, which ensures that the results which follow arise from reputation tracking and not repeated interactions between pairs of individuals. At the end of the generation, payoffs are totaled for all individuals, and strategies reproduce by a replicator dynamic, which may be either biological reproduction or success-biased imitation (see Gintis, 2000). The exact details of how these strategies are transmitted in human societies could change some of the conclusions of the model, however, and so I intend this analysis as an examination of the basic logic of a strategy that tracks and uses reputations.

Imagine now a strategy named tough, which uses the reputation of its opponents to help decide whether to fight or not. Tough prefers to

TABLE 1
Payoffs in the hawk–dove (or chicken) game. Payoffs written are to player on side

Focal individual	Opponent	
	Hawk	Dove
Hawk	$\frac{v-c}{2}$	v
Dove	0	$\frac{v}{2}$

play like a dove (to resolve the conflict peacefully), however a tough individual will play like a hawk (i.e. fight) whenever (1) her opponent escalates (plays as a hawk) or (2) her opponent is in "bad" standing. Tough individuals keep track of player standing such that participating in an escalated contest produces a "good" standing for both individuals involved. Playing like a dove when one's opponent plays like a hawk produces a "bad" standing for the player behaving as a dove. In all other cases, an individual's standing remains unchanged from the previous time period. Table 2 summarizes how interactions affect standing.

I assume that each individual—through observation or gossip—knows the standings of all other players, but that it is known with some error, such that individuals in good standing are mistaken for individuals in bad standing e of the time (simulations indicate that allowing the symmetrical kind of error, while complicating the mathematics, does not substantially alter the dynamics of the model). Let $S_n(A)$ represent the proportion of individuals of strategy A in good standing in time period n. Let p be the frequency of hawks in the population and q be the frequency of tough individuals in the population. Under these assumptions, it is possible to write expressions for the proportion of hawks, doves and tough in "good" standing for any round n.

$$S_n(H) = p + q + (1 - p - q)S_{n-1}(H), \quad (1a)$$

$$\begin{aligned} S_n(D) &= (1 - p - q)S_{n-1}(D) + q(1 - e)S_{n-1}(D) \\ &= (1 - p - qe)S_{n-1}(D), \end{aligned} \quad (1b)$$

$$S_n(T) = p + q + (1 - p - q)S_{n-1}(T). \quad (1c)$$

TABLE 2
Possible behavior interactions and their effects on individual standings

Focal individual's behavior	Opponent's behavior	Focal individual's standing
H	H	Good
H	D	Unchanged
D	D	Unchanged
D	H	Bad

While neither hawks nor tough individuals can enter bad standing by their actions, they can begin a generation in bad standing. Recursions (1a) and (1c) thus define the rate at which both gain good standing. Doves in contrast lose good standing by a constant factor each round.

We will assume for now that everyone begins in good standing, and, since hawks and tough individuals have no way to enter bad standing (they will never back down from a fight), every hawk and tough is always in "good" standing $S_n(H) = S_n(T) = 1$). The standing of doves is reduced each round by a factor $(1 - p - qe)$, and so the proportion of doves in good standing in any round n is

$$S_n(D) = (1 - p - qe)^{n-1}. \quad (2)$$

Using the expressions for the standings of the strategies, it is possible to write per round payoffs for each strategy. In any round n, assuming everyone begins in good standing, the payoffs to each strategy are

$$V_n(H) = (p + q)\frac{v - c}{2} + (1 - p - q)v, \quad (3a)$$

$$V_n(D) = (1 - p - q)\frac{v}{2} + q(1 - e)(1 - p - qe)^{n-1}\frac{v}{2}, \quad (3b)$$

$$\begin{aligned} V_n(T) = {} & p\left(\frac{v - c}{2}\right) + (1 - p - q) \\ & \times \left[(1 - e)(1 - p - qe)^{n-1}\frac{v}{2}\right. \\ & \left. + (1 - (1 - e)(1 - p - qe)^{n-1})v\right] \\ & + q\left[(1 - e)^2\frac{v}{2} + (1 - (1 - e)^2)\frac{v - c}{2}\right]. \end{aligned} \quad (3c)$$

Using expressions (3a–c) and given a chance w that an individual goes on to another interaction, it is possible to write expressions for the expected payoffs to each strategy over the entire sequence of interactions (explained in Appendix A).

$$W(H) = \left((p + q)\frac{v - c}{2} + (1 - p - q)v\right)\left(\frac{1}{1 - w}\right), \quad (4a)$$

$$W(D) = \frac{v}{2}\left[\frac{1-p-q}{1-w} + q(1-e)\left(\frac{1}{1-w(1-p-qe)}\right)\right], \quad (4b)$$

$$W(T) = p\frac{1}{1-w}\left[\frac{v-c}{2}\right] + q\frac{1}{1-w}\left[\frac{v}{2} - \frac{c}{2}(1-(1-e)^2)\right] + (1-p-q)\left[(1-e)\frac{v}{2}, \left(1 - \frac{w(1-p-qe)}{1-w(1-p-qe)}\right) + v\left(e + \frac{w}{1-w}\right)\right]. \quad (4c)$$

ESS Analysis

Using the per-generation payoffs above, the conditions under which tough is an ESS can be derived. Here I will ask first how tough does against each pure strategy, and then I will ask how tough does against the likely mixed equilibrium of hawks and doves ($p = v/c$). Throughout, I assume $0 < v < c$, since when $v > c$ it always pays to play hawk.

HAWK CANNOT INVADE TOUGH

When tough is common, $p \approx 0$ and $q \approx 1$. Using the expressions for $W(H)$ and $W(T)$, hawks can invade an equilibrium of tough whenever

$$e > 1. \quad (5)$$

Since e is a proportion, the above is never satisfied, and hawks cannot invade a population of tough.

TOUGH INVADES HAWK

There is no selection against tough when hawk is common, however if tough ever meets another tough, they will do slightly better than the hawks, since they will recognize each other's good standings and share the resource. Therefore selection against hawks initially increases as the system moves toward the pure tough equilibrium. Thus drift processes will be sufficient to destabilize hawks. Because hawks cannot invade tough [expression (5)] and the system is unstable in the region around $p = 1$, there is no mixed equilibrium of hawks and tough individuals.

TOUGH INVADES DOVE, WHEN $e > 0$

A rare tough individual can invade a population of doves when $W(T) > W(D)$, which simplifies to

$$e > 0, \quad (6)$$

when $p \approx q \approx 0$. When reputations are known without error ($e = 0$), any mix of doves and tough individuals is an equilibrium. All doves start in good standing and tough never provokes a fight with someone in good standing, therefore dove has the same fitness as tough and can drift into the population. The same argument will hold for any mix of dove and tough: in the absence of hawk, a dove holds onto his good standing forever and enjoys the same payoff as tough. A symmetrical argument holds for a rare tough in a population of doves. Thus when $e = 0$, drift and other processes will determine the mix of doves and tough individuals.

However, if tough individuals ever mistake individuals in good standing for individuals in bad standing (or ever otherwise pick fights with individuals in good standing), then they can invade a population of doves. This results from errors flushing out hidden doves and establishing bad standings for them. Once a dove is in bad standing, he never regains good standing. Tough individuals then exploit the dove in all subsequent interactions.

DOVE INVADES TOUGH WHEN ERRORS ARE COMMON

When e is large, tough individuals pay additional costs by getting into a large number of escalated contests with one another. Using expressions (4b,c), rare doves can invade a population of tough when either w is small or e

is large:

$$(2-e)[1-w(1-e)]>\frac{v}{c}. \qquad (7)$$

When w is small, rare doves may not be discovered to be doves before interactions end. Therefore, small numbers of interactions help doves invade. When e is large, tough does worse because the costs of routinely engaging in fights with one another outweigh the advantages of detecting rare doves. For example, when $e=1$, condition (7) is always satisfied. Finally, the ratio v/c encourages doves to invade when it is small, when escalated contests are very costly relative to the value of the resource. The equilibrium proportion of doves reached when both eqns (6) and (7) are satisfied is a complicated and difficult to interpret expression which provides only a small amount of additional insight. I describe it in Appendix B.

TOUGH CANNOT INVADE THE MIXED HAWK/DOVE EQUILIBRIUM, BUT IT OFTEN HAS A LARGE DOMAIN OF ATTRACTION

Analysing the pure equilibria provides insights about the conditions under which tough does better than either pure strategy. However, neither hawks nor doves will exist in isolation when $0<v<c$. (Recall, if $v>c$, the world will be all hawks.) Therefore invading tough individuals will instead encounter the equilibrium mix of hawks and doves, where the frequency of hawks is equal to v/c (doves exist at frequency $1-v/c$).

At the mixed hawk/dove equilibrium, a rare tough individual will invade when $v>c$, which is never true given the assumption $0<v<c$. There is, however, an unstable internal equilibrium of all three strategies which is near the hawk/dove equilibrium when w is large. This internal unstable equilibrium lies at

$$\hat{p}=1-\left(1-\frac{v}{c}\right)\frac{(1-w)+(1-e)}{w(1-e)(2-w)}, \qquad (8a)$$

$$\hat{q}=\left(1-\frac{v}{c}\right)\frac{(1-w)}{w(1-e)(2-w)}. \qquad (8b)$$

When $w=1$ (infinite interactions per generation), this equilibrium lies at $p=v/c$ and $q=0$, which is the mixed hawk/dove equilibrium. When there are many interactions per generation, small perturbations away from the hawk/dove mixed equilibrium are sufficient for tough to invade. For $w=0.95$ (20 interactions per time period, on average), the unstable equilibrium lies very close to the hawk/dove equilibrium, and the domain of attraction for the tough/dove equilibrium is very large. Thus small amounts of migration or non-random interaction will allow tough to invade under those circumstances. When $w=0.5$ (2 interactions on average), the domain of attraction for the hawk/dove equilibrium is larger, and many starting conditions will lead the system to the mixed hawk/dove equilibrium instead of the tough/dove equilibrium.

HAWKS SOMETIMES INVADE THE DOVE/TOUGH MIXED EQUILIBRIUM

It is easy to see that when the right side of expression (8b) >1, the only stable equilibrium will be the mixed hawk/dove equilibrium at $\hat{p}=v/c$ and $\hat{q}=0$. Tough behaves much like a pure retaliator strategy, in the absence of errors and when initial standings are all good. However, the additional advantage of recognizing and exploiting doves allows tough to keep pure hawks out of the population, but only as long as interactions are not too few per generation and errors are not too common. When w is very small, even if errors are quite rare, tough individuals suffer relative to doves, since tough individuals engage in costly fights but lack sufficient interactions to reap the rewards of exploiting exposed doves. This result will be very sensitive to assumptions about the initial standings of individuals, however.

Summary of the System Dynamics

Figure 1 shows representative dynamics for the system with two equilibria. The triangular graph plots the frequencies of the three strategies—hawk, dove and tough—such that each corner represents a pure strategy, a point on a side a mix of two strategies, and each internal point a mix of all three. The lines then show system dynamics from given starting mixtures

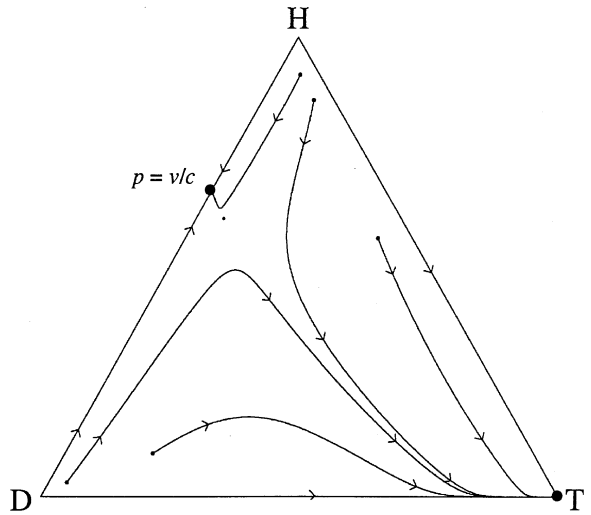

FIG. 1. System with two stable equilibria. The first is the hawk/dove mixed equilibrium at $p = v/c$ and $q = 0$. The second is the pure Tough equilibrium at $q = 1$. $w = 0.85$, $v = 2$, $c = 3$, $e = 0.05$.

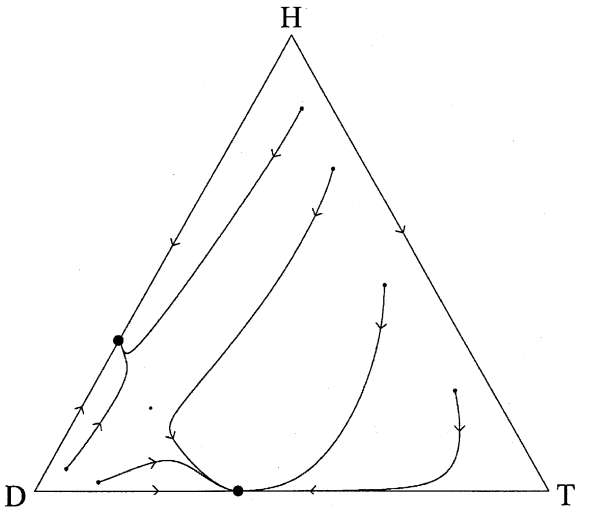

FIG. 2. System with stable equilibria at $p = v/c$ and $q = 0$ and at the mixed dove/tough equilibrium that arises when expression (7) is satisfied and the right- and side of expression (8b) < 1. $w = 0.85$, $v = 1$, $c = 3$, $e = 0.25$.

(the small black dots) of hawks, doves and tough individuals. Thus a line projects from a starting mixture to one of two equilibria, which are represented by larger dots. The lone small internal dot locates the internal unstable equilibrium of all three strategies. In the presence of rare errors ($e = 0.05$) and for 6.66 interactions per generation on average ($w = 0.85$), a small domain of attraction exists for the mixed hawk/dove equilibrium at $p = v/c$ and $q = 0$. A larger domain of attraction exists for the pure tough equilibrium ($q = 1$) in the lower-right corner. Hawks are quickly eliminated within this basin of attraction, and then rare errors slowly reveal and select against doves until the entire population is composed of tough individuals. For larger values of w, the basin of attraction for the mixed hawk/dove equilibrium all but vanishes.

Figure 2 plots the dynamics for a case when a mix of doves and tough individuals is stable. Common mistakes in knowledge of reputation ($e = 0.25$) and a low value of the resource ($v/c = 1/3$) prevent tough from resisting invasion by rare doves, since accidental escalated contests among pairs of tough individuals hurt the strategy. This continues until doves are common enough that the benefits of exploiting them balances the costs of accidental fighting among themselves. Note that the lower resource value relative to the cost of fighting ($v/c = 1/3$ here instead of 2/3 as in Fig. 1) also slightly increases the domain of attraction for the mixed hawk/dove equilibrium, but principally increases the proportion of doves at the mixed dove/tough equilibrium.

Finally, Fig. 3 plots the dynamics for the same two equilibria, but for when mistakes in reputation are very common ($e = 0.5$) and only 2 interactions per generation on average ($w = 0.5$).

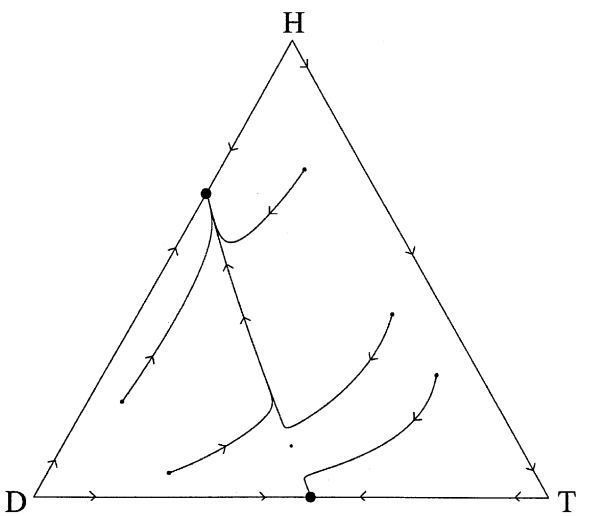

FIG. 3. System with the same equilibria as in Fig. 2, but with high error rates ($e = 0.5$) and a small probability of continuing interaction ($w = 0.5$). $v = 2$, $c = 3$. Under these conditions, the basin of attraction for the dove/tough equilibrium is very small.

The basin of attraction for the mixed hawk/dove equilibrium now dominates, but the mixed dove/tough equilibrium is still stable.

Assuming Individuals are in Bad Standing Aids the Evolution of Tough Buts Hurts it at Equilibrium

If all individuals begin each generation in bad standing rather than good, there are more conditions which lead tough to high frequency. To see why, consider the new expressions for the standings of each strategy in round n, given that everyone begins each generation in bad standing:

$$S_n(H) = p + q + (1 - p - q)S_{n-1}(H), \quad (9a)$$

$$S_n(D) = 0, \quad (9b)$$

$$S_n(T) = p + q + (1 - p - q)S_{n-1}(T). \quad (9c)$$

Hawks and tough individuals gain good standing at a rate proportional to the frequency of hawks and tough in the population, while doves never gain good standing. Using expressions (3a–c), with appropriate changes, and eqns (9a–c), the per generation payoffs to each strategy become

$$W(H) = \frac{1}{1-w}\left[(p+q)\,\frac{v-c}{2} + (1-p-q)v\right], \quad (10a)$$

$$W(D) = \frac{1}{1-w}\left[(1-p-q)\,\frac{v}{2}\right], \quad (10b)$$

$$W(T) = \frac{1}{1-w}\left[p\,\frac{v-c}{2} + (1-p-q)v\right] + q\,\frac{v-c}{2} + wq\left[(1-e)^2 S_2(T)^2\,\frac{v}{2} + (1-(1-e)^2 S_2(T)^2)\frac{v-c}{2}\right] + w^2 q[\ldots] + \cdots. \quad (10c)$$

From expressions (10b,c), doves can now invade a population of tough individuals when

$$\frac{v}{c} < \frac{1 - w + e(2-e)}{2 - w}. \quad (11)$$

and a mixed equilibrium similar to before exists. However, the mixed hawk/dove equilibrium is no longer stable against invading tough individuals. The payoff to a tough against a dove is the same as the payoff to a hawk against a dove, i.e. $V(T|D) = V(H|D)$. Likewise, the payoff to a tough against a hawk is the same as the payoff to a hawk against a hawk: $V(T|H) = V(H|H)$. Therefore the only payoffs differentiating hawk and tough is how tough does against tough compared to how hawk does against tough. In any round in which any tough individual is in good standing,

$$V(T|T) > V(H|T). \quad (12)$$

Thus if interactions ever continue past the first round ($w > 0$), a small number of tough individuals can drift in and invade the mixed hawk/dove equilibrium. Whether tough also dominates dove depends upon the values of e, w and v/c, as in expression (11). But since tough will always do better than hawk, no interior equilibrium exists. Figure 4 plots the dynamics of the system with

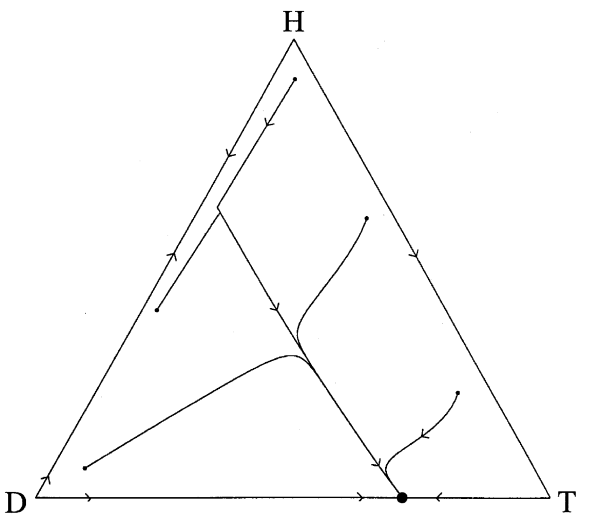

FIG. 4. System dynamics with same parameter values as Fig. 3 ($w = 0.5$, $e = 0.5$, $v/c = 2/3$) but with the assumption that all individuals begin each generation in bad standing. With this assumption, only the dove/tough equilibrium is stable.

the same parameters as Fig. 3 ($w = 0.5$, $e = 0.5$, $v/c = 2/3$). The system now has only one stable equilibrium, along the dove/tough line. It is important to note, however, that selection in the region of $p = v/c$, $q = 0$ is very weak.

I have just argued that assuming that unknown individuals are in bad standing allows tough to invade. However, once at equilibrium, the assumption that any unknown individuals are in bad standing is costly for the tough strategy. To see why, consider a population of tough individuals who assume a proportion s of unknown individuals are initially in good standing. I will denote this strategy with T_s. The payoff to such a suspicious tough strategy, when it alone exists at equilibrium, is

$$W(T_s) = (1-e)^2(1-s)^2\frac{v}{2} + (1-(1-e)^2(1-s)^2)\frac{v-c}{2} + \frac{w}{1-w}\left[(1-e)^2\frac{v}{2} + (1-(1-e)^2)\frac{v-c}{2}\right]. \quad (13)$$

In the first round, any interaction in which either party commits an error or assumes the other to be in bad standing leads to an escalated contest, but these contests also establish everyone in good standing for all following rounds. Each individual then enjoys $v/2$ until interactions end, discounted by accidental fights brought about by errors in reputation (e). Now consider an invading tough strategy $T_{s+}\delta$, which assumes a proportion $s+\delta$ of individuals are in good standing initially. The fitness of such a strategy, when T_s is common, is

$$W(T_{S+\delta}) = (1-e)^2(1-s)(1-s-\delta)\frac{v}{2} + (1-(1-e)^2(1-s)(1-s-\delta))\frac{v-c}{2} + \frac{w}{1-w}\left[(1-e)^2\frac{v}{2} + (1-(1-e)^2)\frac{v-c}{2}\right]. \quad (14)$$

From expressions (13) and (14), $T_{s+\delta}$ will invade T_s whenever $\delta > 0$. Thus payoff sensitive forces will act differently on assumptions about initial standing, depending upon the frequency of tough in the population. If tough individuals remain common long enough, initial assumptions about standing should become almost entirely positive. However, the presence of doves at equilibrium or migration in a structured population would likely sustain an intermediate value of s.

Discussion

Under a wide range of conditions, the tough strategy comes to dominate the population. When communities are stable such that reputations matter and reputations are not known without error but known well, tough either excludes hawks and doves or exists at a high frequency among dove individuals. In this section, I discuss some additional modeling concerns. Then I turn to interpreting the model in light of our ethnographic understanding of existing honor systems.

IMPERFECT KNOWLEDGE PRODUCES A VERY SIMILAR MODEL

I have also simulated a model in which there is a chance k a tough individual knows the standing of her opponent. In the event she does not know her opponent's standing, s of the time she assumes the opponent to be in good standing. The rest of the time, she assumes the opponent to be in bad standing and behaves accordingly.

This model generates behavior and equilibria very similar to the model with errors. For that reason, I will not discuss it in detail here. It is easy to see, however, why the two models are so similar. Both systems flush out doves who have not yet been discovered and cause costly escalated contests between pairs of tough individuals. In the model with imperfect knowledge, the errors (e) of the first model are some combination of knowledge (k) and the chance of assuming an opponent to be in good standing (s).

An interesting conceptual difference, however, is that the chance an individual would know the reputation of a random individual declines as community size increases. Thus we might interpret either e or $1-k$ as increasing with the size and anonymity of communities. These

communities are then more likely to occupy mixed dove/tough equilibria than small personal communities, which would be attracted to pure tough equilibria. Simultaneously, however, the domain of attraction for the dove/tough equilibrium shrinks as k decreases or e increases, meaning that it would be very hard for tough to get started.

Additional Strategies: *Reputation Seeking*

A number of additional strategies could be proposed in this environment. I have analysed the strategies which appear both most sensible and most illustrative. However, one interesting amendment to make to the tough strategy is to allow it to take its own standing into account. Sugden's (1986) standing strategy, Contrite Tit-for-Tat (Boyd, 1989), cooperates whenever it has bad standing. This allows it to quickly regain good standing in the event of performance errors. Tough would benefit from behaving in a similar way, playing hawk whenever it is in bad standing. The analysis for when $S_1(T) = 1$ would remain unchanged. However, whenever some individuals begin each generation in bad standing ($S_1(T) < 1$), this change would lead tough individuals to gain good standing more quickly, increasing their expected payoffs relative to the strategy I analysed above. Additionally, this process of seeking good standing will flush out doves more quickly. For these reasons, the analyses above could be considered to be stacked against tough.

However, once we allow tough to consider its own standing in making decisions, we must also allow for errors in knowledge of its own standing. Errors of this kind will lead to reduced expected payoffs for tough individuals. Before, two tough individuals shared the resource with probability $S_n(T)^2(1-e)^2$. With the modified strategy, and if either can incorrectly believe himself to be in bad standing with probability f, the chance they share the resource is reduced to

$$S_n(T)^2(1-e)^2(1-f)^2. \qquad (15)$$

If either is in bad standing, he will provoke a contest. If either believes the other to be in bad standing or believes himself to be in bad standing, he will likewise provoke a contest. Only when none of these errors occur and both are in good standing will they share the resource.

Simulation and analysis shows that this amended tough strategy changes none of the qualitative results derived before, unless f is large. When f is very large, the domains of attraction for tough shrink considerably, exactly as e affects the domains of attraction and equilibria. It is easy to see from expression (15) how f has the same form of effect as e. When f is not very large though, tough individuals gain good standing quickly enough that additional escalated fights between tough individuals have little impact on the long-run payoffs. There is good reason to believe that f should be much lower than e and usually very close to zero, given that an individual is likely to have more information about his own standing than the standings of others.

Cultures of Honor

The tough strategy succeeds, when it does, because it makes life costly for hawks *as well as* doves. Unlike Retaliator (Maynard Smith, 1982), which plays dove unless provoked, tough also uses information from how others have played in the past in order to take advantage of doves. In this paper, I have derived expressions which shed light on the conditions under which such a strategy does well, and I have shown that a reputation-based strategy can often succeed in regulating conflicts. Beyond verifying the logical consistency of an argument, there are two key things models of this sort can do for us. First, they can provide counter-intuitive results which stimulate tests and elaborations of theory. Second, they can provide qualitative predictions which allow us to directly examine the power of the model as an explanatory tool. One common way in which this is done is by "comparative statics," in which we attempt to match the equilibrium outcomes of the model under higher and lower values of key parameters with data. The goal is to see if, for example, more stable and less anonymous communities more commonly have cultures of honor. Here I comment on the comparative statics of the key parameters

in the model and offer intuitions about the real-world relevance of those parameters.

Unsurprisingly, the number of interactions ($1/(1-w)$, the stability of communities) has a positive effect on the viability of the tough strategy. Any strategy which uses past behavior as a guide depends upon stable communities. The same kind of result is typical in iterated prisoner's dilemma games, as well (Axelrod & Hamilton, 1981; Trivers, 1971). The interpretation of "long" and "stable" is subjective, of course. In these models, even $w = 0.8$ (5 interactions on average per generation) strongly favors tough under most conditions. Most human communities are substantially more stable than that. This suggests that the dynamics of reputation play out over short enough time periods such that strategies which track reputation can accrue benefits in the medium run as well as the long run. The recursions for standings I derived previously imply this result, as reputation dynamics are exponential in most cases. This means the standings of the different strategies will converge to their equilibria very rapidly. Numerical work confirms this intuition.

The ratio of the value of the resource (v) to the cost of losing a fight (c) strongly affects the viability of the tough strategy, as well. When v is large relative to c, tough individuals do well compared to other strategies. To see why this is true, compare the payoffs of a dove and a tough individual when the two types are mixed in a population. Doves always prudently avoid escalated contests, so they never suffer the cost c, but they also never claim the entire resource, v. Instead, they either share the resource or give it up entirely. Tough individuals pay the price of entering into occasional escalated contests, so they suffer the cost of losing, c. But tough individuals also take advantage of doves, once doves earn their reputation as doves, and claim the entire value of the resource, v, whenever they encounter one. The balance between occasional escalated contests and being able to exploit doves is what drives the comparative advantage of tough over dove. When fights are more costly (c is larger relative to v), the price tough individuals pay for entering into them with one another begins to outweigh the advantage accrued by taking advantage of doves. In the model, when v is one-eighth as large as c ($w = 0.95$ and $e = 0.05$), 60% of the population is dove at equilibrium, and only 40% tough individuals. When v is a sixteenth of c, almost no tough individuals are present at equilibrium. For large values of v relative to c, the equilibrium proportion of tough approaches one.

Ethnographically, we might consider the value of the resource as analogous to the consideration of stakes Nisbett & Cohen (1996) suggest drives the evolution of cultures of honor. In herding economies, for example, the resource under contest in a family's herd. This herd is very valuable in terms of a household's subsistence and status. For farmers, on the other hand, any resource under contention is not likely to be as large. Crops cannot be easily stolen in large amounts. We might regard the cost of losing a contest as quite high, however, given that firearms, iron weapons and poisons dramatically raise the costs of contests in most human societies.

The rate of errors in assessing standing, e, can be interpreted as both a feature of the social environment or as a feature of the tough strategy. As a feature of the environment, e (or k and s in the limited knowledge model) reflects the familiarity of the members of the community and the observability of the outcome of contests. It will also incorporate the degree of gossip and informal dissemination of knowledge about these outcomes. It seems unlikely that many human societies would have very high values of e, due to limited knowledge. Even in settings such as African herding economies, where households are very scattered and mobile, people have a great deal of knowledge about one another. High values of e due to knowledge might characterize new and informal communities, such as novel mining or frontier towns, however.

Knowledge of reputation may not be such a key parameter, however. Note that in general the tough strategy is extremely robust to errors in reputation. When w is large, even with $e = 0.75$, the domain of attraction for the dove/tough equilibrium is very large [expression (8b)], and tough individuals exist at high frequency. This is in contrast to indirect reciprocity strategies,

which use reputation to direct altruism and are very sensitive to errors in knowledge (Panchanathan & Boyd, under review).

Is Taking Advantage of Doves "Honorable"?

A few readers of an earlier version of this paper expressed the intuition that "honorable" individuals, unlike tough individuals attacking doves, would eschew fighting with weak or dishonorable individuals. The data behind this aspect of the tough strategy is the widespread perception by individuals in cultures of honor that reputation is valuable for protection from others like onself (Nisbett & Cohen, 1996). Lawless regions are not necessarily filled with do-gooders who defend their pride, but with opportunists and bullies making their way among many other opportunists. Nisbett and Cohen draw several persuasive analogies between many classic cultures of honor and inner-city gang culture, where reputation may serve as both protection and earn one access to resources. Additionally, tough individuals do not necessarily fight with doves, but simply exploit them by seizing resources, either through coercion or custom. At the level of abstraction of the models, there is no difference.

I take very seriously the alternative possibility, however, that the ideal of doing right in the world but defending oneself at all costs accurately describes several ethnographic and historical settings. Not every honor society need be the Mafia. It may be that a mixed population of retaliators and individuals who use reputation information to exploit doves generates the value of a truly honorable strategy. For example, Fischer (1989) describes the back-country environments and their gangs of "banditti" which stimulated vigilante justice in 18th century British America. In the next section, I suggest some ideas for modeling more complicated strategy sets of this kind, which might result from behavior contingent upon differences in resource holding power. Fleisher's (2000) recent ethnography of cattle raiding in northern Tanzania, however, strongly suggests that in other contexts the honorable and the exploiting are indeed one and the same.

Further Problems: Multiple Contests, Bluffing and Asymmetries

The model in this paper indicates that reputation can successfully regulate contests, but not under all conditions. The model could be extended in a number of useful ways. One key feature of the culture of honor argument missing from the models I have analysed here is the multiple contest nature of reputation. That is, people supposedly use the outcome of trivial contests as signals of how people will behave in non-trivial matters. This model includes only one type of contest. A model in which players interact in a trivial contest t of the time and a high-stakes contest $1-t$ of the time would allow one to explore how the frequency of low- and high-stakes contests affects the tough strategy. Additionally, it allows for the invasion of strategies which act like tough in the low-stakes contests but play dove in the high-stakes ones. Such bluffing strategies might destabilize tough under some conditions and shed more light on what factors influence the viability of honorable strategies. Finally, adding asymmetries in fighting ability would lead to an exploration of the conditions under which weaker individuals would be willing to fight just for the reputation value, even when they are likely to loose.

I thank Rob Boyd, Joseph Henrich, Aimee Plourde, two thoughtful anonymous reviewers, and the herders and farmers of Usangu, Tanzania for helpful comments and advice on both the models and earlier versions of this paper.

REFERENCES

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation in biological systems. *Science* **211,** 1390–1396.

Boyd, R. (1989). Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *J. theor. Biol.* **136,** 47–56.

Fischer, D. (1989). *Albion's Seed: Four British Folkways in Americ*a. New York: Oxford University Press.

Fleisher, M. L. (2000). *Kuria Cattle Raiders*. Ann Arbor: University of Michigan Press.

Gintis, H. (2000). *Game Theory Evolving*. Princeton: Princeton University Press.

Johnstone, R. A. (2001). Eavesdropping and animal conflict. *Proc. Natl. Acad. Sci.* **98,** 9177–9180.

LEIMAR, O. & HAMMERSTEIN, P. (2001). Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. London B* **268,** 745–753.
MAYNARD SMITH, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
MAYNARD SMITH, J. & Price, G. R. (1973). The logic of animal conflict. *Nature* **146,** 15–18.
MCELREATH, R. (2001). Chapter 3: Lex Talionis: Reputation can regulate conflict in Hawk–Dove contests. Ph.D., University of California, Los Angeles.
NISBETT, R. E. & COHEN, D. (1996). *Culture of Honor: The Psychology of Violence in the South*. Boulder: Westview Press.
NOWAK, M. & SIGMUND, K. (1998a). Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573-577.
NOWAK, M. A. & SIGMUND, K. (1998b). The dynamics of indirect reciprocity. *J. theor. Biol.* **194,** 561–574.
PANCHANATHAN, K. & BOYD, R. (under review). A tale of two defectors: the importance of standing in the evolution of indirect reciprocity. *J. theor. Biol.*
SUGDEN, R. (1986). *The Economics of Rights, Co-operation, and Welfare*. New York: Blackwell.
TRIVERS, R. (1971). The evolution of reciprocal altruism. *Quar. J. Biol.* **46,** 35–57.

APPENDIX A

Using expressions (3a–c), it is possible to write closed expressions for the per-generation payoffs to each strategy. Let w be the chance that an individual goes on to interact with another random individual in the population. Then the expected number of interactions is $1/(1-w)$. The payoff to a hawk is just $V_1(H)$ multiplied by this expectation.

The payoff to a dove is more complicated. For simplicity of notation, let $x = 1-p-qe$.

$$W(D) = V_1(D) + wV_2(D) + w^2V_3(D) + \cdots$$

$$= V_1(D) + w\left\{(1-p-q)\frac{v}{2} + q(1-e)x\frac{v}{2}\right\} + w^2\left\{(1-p-q)\frac{v}{2} + q(1-e)x^2\frac{v}{2}\right\} + \cdots$$

$$= V_1(D) + \frac{v}{2}(1-p-q)\ \{w + w^2 + w^3 + \cdots\} + \frac{v}{2}q(1-e)\{wx + w^2x^2 + w^3x^3 + \cdots\}. \qquad (1b)$$

Closing the infinite sums:

$$W(D) = V_1(D) + \frac{v}{2}(1-p-q)\left\{\frac{w}{1-w}\right\} + \frac{v}{2}q(1-e)\left\{\frac{wx}{1-wx}\right\}.$$

The above expression readily yields expression (4b) in the text. The payoff to a tough individual is derived similarly. Since interactions with hawks and other tough individuals are not affected by standing (but only by errors), these portions of the payoff can be closed straight off, by using the expected number of interactions $1/(1-w)$:

$$W(T) = p\frac{1}{1-w}\left\{\frac{v-c}{2}\right\} + q\frac{1}{1-w}\left\{\frac{v}{2} - \frac{c}{2}(1-(1-e)^2)\right\}$$

$$+ (1-p-q)\left\{\begin{array}{c}(1-e)\frac{v}{2} + ev + w\left((1-e)x\frac{v}{2} + (1-(1-e)x)v\right)\\ +w^2\left((1-e)x^2\frac{v}{2} + (1-(1-e)x^2)v\right) + \cdots\end{array}\right\}.$$

The term for dove interactions is then closed by splitting the terms within the infinite series and closing each as usual

$$W(T) = p\frac{1}{1-w}\left\{\frac{v-c}{2}\right\} + q\frac{1}{1-w}\left\{\frac{v}{2} - \frac{c}{2}(1-(1-e)^2)\right\}$$

$$+(1-p-q)\left\{\begin{array}{c} ev+(1-e)\,\dfrac{v}{2}(1+wx+w^2x^2+\cdots) \\ +v(w+w^2+\cdots)-(1-e)v(wx+w^2x^2+\cdots) \end{array}\right\}$$

$$=p\,\frac{1}{1-w}\left\{\frac{v-c}{2}\right\}+q\,\frac{1}{1-w}\left\{\frac{v}{2}-\frac{c}{2}\,(1-(1-e)^2)\right\}$$

$$+(1-p-q)\left\{ev+(1-e)\,\frac{v}{2}\left(\frac{1}{1-wx}\right)+v\,\frac{w}{1-w}-(1-e)\,v\,\frac{wx}{1-wx}\right\}.$$

After some rearranging, the above yields expression (4c) in the text.

APPENDIX B

When expressions (6) and (7) are satisfied, and eqn (8b) <1, there is an equilibrium mix of dove and tough individuals. This equilibrium is found by assuming $p=0$, setting $W(T)=W(D)$, and solving for q. The resulting expression is quadratic, with only one root falling between zero and one:

$$\hat{q}=\frac{wv-(1-w)(2-e)c+\sqrt{(wv)^2+(1-w)(2-e)c\{(1-w)(2-e)c-2wv(1-2e)\}}}{2we(2-e)c}. \quad \text{(A.1)}$$

This expression is messy and hard to interpret. A plot of the above as a function of e and w, however, is revealing (Fig. A1).

For low values of w, increases in errors, e, lead to more tough at equilibrium. This is because errors help flush out hidden doves, which is important when there are few interactions per generation. The faster tough exposes these doves, the more it is able to exploit them. For high values of w, however, errors decrease the equilibrium proportion of tough. This results from the long-term costs of tough individuals fighting with one another, long after all doves have been placed in bad standing.

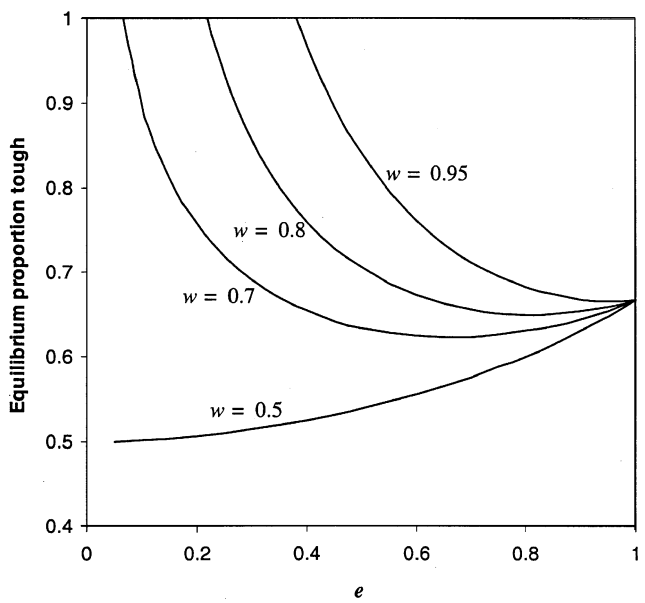

FIG. A1. Plot of expression (A.1) against e, for $v/c=2/3$ and four values of w.