

the absence of a definitive historical record, therefore, the empirical strength of the current paper rests entirely on the charm, as a metaphor for human interactions, of the random matching one-shot Prisoner's Dilemma model with imperfect identification of types. But this is not the only charming model available, and some of the others (including the same one with different parameter values) do not admit a stable proportion of cheaters, or do not admit cheaters at all. Also, if taken seriously, the model chosen makes predictions about sociopathy that do not seem to be true.

I will address these criticisms in reverse order. Recall that if a mixed strategy profile is evolutionarily stable, then if a particular strategy A increases (or decreases) in frequency, the payoffs to all the strategies must change so that A is relatively disadvantaged (or advantaged). Natural selection will then ensure that strategy A goes back to its proper place. If we are to take the current application of Frank's (1988) model seriously, then the proportion of sociopaths in a society should follow a similar dynamic.

But does it? Would primary sociopaths in an ancient society with "too many" of their own kind have had difficulty gaining access to resources and mating partners? We just do not know, and the target article is not helpful. Perhaps cooperators would become more diligent, but if cheaters ganged up under a charismatic Attila one suspects large numbers would be an advantage.

The dynamic for secondary sociopathy is discussed in the paper, but things seem to go the wrong way. One has to be careful here – if I move from Kingston to New York City and as a result my kids are more likely to become sociopathic, this could be because environmental differences lead New York to have a higher proportion of cheaters. Rather, suppose that a small group of young sociopaths move to Kingston, all else the same. According to the Frank/Mealey model, Kingston children will now be less likely to become sociopathic. I have no evidence; but like most parents, I think not.

Frank's (1988) model is a variant of the round robin, infinitely repeated Prisoner's Dilemma introduced by Axelrod (1984) in his celebrated competition. But there are other models that have equal "charm" and are therefore equally (un)likely to capture the essence of the conditions facing primitive homo sapiens. Here are two examples:

Consider a model where individuals match up randomly, play a one-shot Prisoner's Dilemma, and then have the choice of continuing or terminating the match (Carmichael & MacLeod 1994; Stanley 1993). If the match ends, both parties go back, anonymously, to the matching market. If not, the partners may stay matched until death, continuing to play a Prisoner's Dilemma each period. For a modern image, think of the matching market as a large, dimly lit singles' bar.

Even though cooperators play a repeated Prisoner's Dilemma, the strategy "tit for tat" does very poorly. It is quickly invaded by cheaters who defect at the first opportunity and then move on to a new match. An interesting evolutionarily stable strategy is for cooperators to offer (and demand) an exchange of gifts at the beginning of any new match (Carmichael & MacLeod 1993). Cheaters in this society would have to buy a succession of gifts, and this effectively screens them out. This model makes quite a few predictions about the form of the gifts that must be used.<sup>1</sup>

Here is another one (Carmichael 1994). Suppose we retain the one-shot matching framework of Frank but change the game from a Prisoner's Dilemma to a bargain. People meet and have to decide how to divide the spoils of some joint venture (the carcass of some animal, perhaps). If they can agree quickly on a division, then all is well. If they cannot, the spoils disappear, dragged away by a hyena.

Strategies that do well in these bargains will proliferate into the future. An intraspecies arms race might develop, where "bargaining ability" grows over time. Rational and unemotional

bargainers will be vulnerable to the "terrible twos" strategy of demanding almost everything, backed up with the emotional threat to ensure that otherwise the hyena gets everything. Faced with such an opponent, a rational bargainer cuts his losses, takes what is offered, and moves on. Of course an entire society of two-year-olds does very poorly, and can be invaded by a group whose members fight if they do not receive at least half the spoils. This strategy is evolutionarily stable – it quickly reaches agreement with itself and can do no worse than any invader it meets.<sup>2</sup> Again, in this simple model, there is no room for cheaters.

Readers will recognize the "bourgeois" strategy of Maynard Smith (1982), but there are some new twists. In particular, if people are of two types, there are many equilibria where one type does better than the other. If men fight whenever they get less than one-third and women fight whenever they get less than two-thirds, for example, this is evolutionarily stable. Equilibria like these require that one's notion of territory be *socially* determined. There must be a way for early experience and teaching to establish and coordinate internal notions of what one deserves out of life. Sociobiology can therefore account for the existence of systemic discrimination, and society may indeed, at least partly, be to blame.

The point, of course, is not that these are better models of reality than the one used in the target article, but they do seem equally plausible, and they have implications that are at least as attractive and intriguing. Perhaps they each capture relevant but separate aspects of reality. If so, Mealey's conclusion – that evolution is unable to rid society of a small proportion of cheaters – is not robust. (The rule of emotions in these models, by contrast, is robust.) Cheaters do prosper, no doubt. But until we have excellent evidence about the exact nature of early human society, or until we can show that in any sensible evolutionary model there will survive a small proportion of cheaters, sociobiology will not be able to tell us why.

#### NOTES

1. Among other things, gifts should have low use value, be overpriced, and should be hard to recycle as gifts in a subsequent match. Cut flowers and chocolates work – house plants and money do not.

2. Unless, of course, the invader has a weapon. The arms race in this model is real.

## Prisoner's Dilemma, Chicken, and mixed-strategy evolutionary equilibria

Andrew M. Colman

Department of Psychology, University of Leicester, Leicester LE1 7RH, England. [amc@leicester.ac.uk](mailto:amc@leicester.ac.uk)

**Abstract:** Mealey's interesting interpretation of sociopathy is based on an inappropriate two-person game model. A multiperson, compound game version of Chicken would be more suitable, because a population engaging in random pairwise interactions with that structure would evolve to an equilibrium in which a fixed proportion of strategic choices was exploitative, antisocial, and risky, as required by Mealey's interpretation.

In a target article of exceptional scholarship and originality, Mealey has put forward an interesting new interpretation of sociopathy. Given the vast range of material covered by the article and the limited space available for my commentary, I shall confine my comments to the specific game theoretic model that underpins Mealey's interpretation. I shall argue that it cannot do what is required of it, and I shall suggest an alternative.

Like most earlier theorists who have used game theory to explain the evolution of social behavior, starting with Maynard

Smith (1974; Maynard Smith & Price 1973), Mealey relied on a two-person game, specifically the Prisoner's Dilemma game. A generalized payoff matrix for any two-person, two-choice game can be represented as follows:

	C	D
C	R	S
D	T	P

The row and column players each choose between two pure strategies, C and D, the payoffs shown in the matrix are those to the row player. Thus the payoff to the row player following a C choice is R or S, depending on whether the column player chooses C or D, respectively, and following a D choice it is T or P, depending on whether the column player chooses C or D, respectively. In the Prisoner's Dilemma game, C represents cooperation and D defection, and by definition  $T > R > P > S$ , so that the row player receives the best payoff by choosing D (defect) while the column player chooses C (cooperate), the second-best payoff by choosing C while the column player chooses C, and so on. Although it is customary to show only the row player's payoffs, the game is the same from the column player's point of view, so that the column player also gets the best payoff by choosing D while the row player chooses C, and so on.

The standard two-person model is of limited value in determining evolutionary processes. We need to establish what will happen in an entire population in which individuals interact with one another in pairwise games with this strategic structure, assuming that the payoffs represent units of *Darwinian fitness* (the lifetime reproductive success of the individual players) and that the propensity to choose C or D is heritable. For this purpose, we need to construct a multiperson *compound game* (Colman 1982, pp. 163–66, 243–49), in which it is assumed that every player plays the same average number of two-person games either with each of the others or with a random sample of the others.

Considering the situation from a single player's viewpoint, suppose that the number of other players is  $n$  and the number of other players choosing C is  $c$ . The total payoff to a player choosing C, denoted by  $P(C)$ , and the total payoff to a player choosing D, denoted by  $P(D)$ , are then defined by the following payoff functions:

$$P(C) = Rc + S(n - c),$$

$$P(D) = Tc + P(n - c).$$

The total payoff to a player adopting a mixed strategy is just a weighted average of  $P(C)$  and  $P(D)$ .

The values of the  $P(C)$  and  $P(D)$  payoff functions at their endpoints are found by setting  $c = 0$  and  $c = n$ . Thus, if none of the other players chooses C (i.e.,  $c = 0$ ), the payoff to a solitary C chooser is  $Sn$  and the payoff to a D chooser is  $Pn$ . If all of the other players choose C (i.e.,  $c = n$ ), then a C chooser gets  $Rn$  and a solitary D chooser is  $Tn$ . It is clear that in the case of the Prisoner's Dilemma game  $Tn$  can be interpreted as the *temptation* to be the sole D chooser,  $Rn$  the *reward* for collective cooperation,  $Pn$  the *punishment* for collective defection, and  $Sn$  the *sucker's payoff* for being the sole C chooser.

Figure 1(a) shows clearly that, in the case of the Prisoner's Dilemma game (with  $T > R > P > S$ ), the  $P(D)$  payoff function strictly dominates the  $P(C)$  payoff function, which means that a D choice pays better than a C choice irrespective of the number of others choosing C. The evolutionarily optimal strategy is therefore not frequency-dependent, and the population will (regrettably) evolve to a stable equilibrium in which every player chooses D in every two-person encounter. This means that the Prisoner's Dilemma game cannot provide a basis for

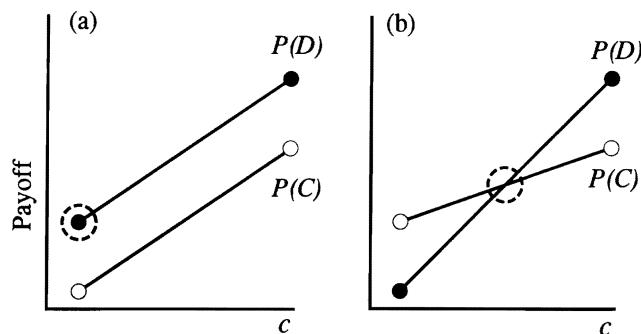


Figure 1 (Colman). Multiperson compound games based on  $2 \times 2$  matrices. Panel (a) on the left is multiperson Prisoner's Dilemma; (b) on the right is multiperson Chicken. The  $P(C)$  and  $P(D)$  functions indicate the payoffs to a player choosing C or D when  $c$  of the other players choose C. Dashed circles indicate stable equilibria.

Mealey's interpretation of sociopathy, in which the "cheater strategy" (the D choice) corresponds to various criminal, delinquent, and generally antisocial or predatory forms of behavior that she claims exist at a low frequency in every society and are maintained through frequency-dependent Darwinian selection.

A more appropriate game theoretic model might be a compound version of the game of Chicken, which Maynard Smith (1976; 1978) and Maynard Smith and Price (1973) call the Hawk-Dove game. This game is defined by the inequalities  $T > R > S > P$ , and the  $P(C)$  and  $P(D)$  payoff functions are shown graphically in Figure 1(b). In this case, the population will evolve to a mixed-strategy equilibrium point, where the two payoff functions intersect. To the left of the intersection, when relatively few of the others choose C ( $c$  is small), the C function lies above the D function, which means that the fitness payoff from a C choice is higher than from a D choice, so the number of C choosers will increase relative to D choosers and the outcome will move to the right as  $c$  increases. To the right of the intersection, exactly the reverse holds: D choosers will increase relative to C choosers and the outcome will move to the left as  $c$  decreases. At the intersection, and only there, the strategies are best against each other and are in equilibrium, and any deviation from the mixture at that point will tend to be self-correcting. By setting the parameters (values of the payoffs  $T$ ,  $R$ ,  $S$ , and  $P$ ) appropriately, the intersection point, and thus the proportion of "predatory" D-choices, can be made as small as required.

It appears, therefore, that the Prisoner's Dilemma game cannot underpin an evolutionary explanation of sociopathic behavior, but that a multiperson compound game version of Chicken, in which cheating is at least frequency-dependent, might be more promising. Chicken is the archetypal *dangerous* game, because a player can outdo a co-player only by cheating (choosing D) while the co-player behaves cautiously (by choosing C), and any such attempt to get the best payoff ( $T$  in the payoff matrix above) involves a necessary risk of the worst possible payoff ( $P$ ). The interpretation of criminal, delinquent, and generally antisocial behavior in terms of strategic choices seems more natural in the game of Chicken. (For a more detailed discussion of the strategic properties of Chicken and some observations on its application to antisocial and criminal behavior, see Colman 1982, pp. 98–104; 1995, sect. 9.6.)

ACKNOWLEDGMENT

Preparation of this commentary was facilitated by Grant No. L122251002 from the Economic and Social Research Council as part of the Framing, Salience and Product Images project.