The Economic Journal, **112** (*July*), 679–703. © Royal Economic Society 2002. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

SOME EXPERIMENTAL EVIDENCE ON THE EVOLUTION OF DISCRIMINATION, CO-OPERATION AND PERCEPTIONS OF FAIRNESS*

Shaun Hargreaves-Heap and Yanis Varoufakis

When two people agree to trade, they unlock a mutual benefit, resolve a potential conflict and gain in proportion to their relative 'aggression', eg the Hawk–Dove game. In an experiment with this game, a discriminatory convention evolved when half of the players were randomly assigned a red and the other half a blue label. Later, the same players were also offered the option of co-operating. Those disadvantaged by the colour-based discriminatory convention co-operated with one another most of the time while the rest did not. The paper offers an explanation of these observations based on a modification of Rabin (1993).

The weaker are always anxious for justice and equality. The strong pay heed to neither.
(Aristotle, *Politics*, s1318b)

Many economic interactions mix mutual benefit with a measure of conflict. For instance, when two people trade, there is often more than one price where both will benefit. The high end of the range favours the seller while the lower advantages the buyer. So, when they settle on a price and trade, they unlock a mutual benefit and resolve a potential conflict. The Hawk–Dove (HD) game captures these elements, albeit in a rather simple way as each player only has a choice between being a hard bargainer (a hawk) and a soft one (a dove). Nevertheless, this is why it is regarded as one of the classic games of social life and why it is important to be able to predict behaviour in this game.

Prediction, however, is difficult in the HD game for reasons that relate to some fundamental issues in game theory. The game has multiple Nash equilibria and the equilibrium selection problem is not readily solved, if we stick with the mathematical description of the game, by an appeal to salience. The symmetrical solution, for instance, echoes the symmetry of the game, but it is not a Nash equilibrium and so does not seem a good candidate for salience. Likewise, the two pure strategy equilibria are symmetrical with one another and so the appeal of one looks as strong as the other.

It is possible, however, that a factor that is extraneous to the mathematical description of the game might make one of these asymmetric equilibria salient. Indeed, evolutionary theorists argue that extraneous factors which distinguish between the players and which are common knowledge can 'seed' conventions which advantage one type of player relative to another, see for example, Sugden (1986) and Weibull (1995) – see also Lewis (1969) on conventions.

* We thank Bob Sugden and an anonymous referee for many valuable comments. Thanks are also due to George Krimpas, for his comments and support, as well as seminar participants at the Universities of Sydney, Valencia, Louvain-la-Neuve and Athens. This project was funded by Australian Research Council grant no. 24657 and Faculty of Economics, University of Sydney grant no. 25663. All errors are ours.

Others find this explanation of equilibrium selection implausible because the inequalities in outcome are supported only by convention and owe nothing to power, ability or principles of fairness, etc. While some evolutionary theorists concede that principles of fairness may play a role in equilibrium selection in such games, they also sometimes follow Hume (1740) and argue that these ideas of fairness themselves develop out of the emerging conventions. Thus, a convention that evolves in the playing of HD may come to be associated with a set of self-validating normative expectations regarding what is fair. These ideas may then come to affect behaviour in other games (Sugden, 1986, 2000). The main purpose of this paper is to see whether these processes of convention and idea formation occur in simple experimental games.

First, we tested for whether a convention emerges in the HD game when players are given a piece of distinguishing extraneous information. In particular, players were given either a red or blue identifying colour in the experiment before playing HD and we tested whether the subsequent behaviour was consistent with people following a convention founded on this initial arbitrary colour assignment. Of course, the distinguishing features that might be used in social life are liable to be more complex in origin than this. Nevertheless it is helpful to know whether conventions can arise in this rather simple experimental setting as it gives an insight into whether the same kind of mechanisms could underpin the generation of conventions in society more generally.

Second, we investigated how ideas of fairness associated with the evolutionary emergence of a convention in one game might affect play in another game. There are two possibilities here. The principle of fairness generated in one game can act as an equilibrium selection device in other games. Alternatively, these ideas of fairness could feed into a new equilibrium concept: that is, the players' concern to be fair may support non-Nash equilibria in these games. For example, it is sometimes argued in behavioural economics that the selection of the co-operative strategy in the prisoners' dilemma game can be explained through the introduction of 'psychological' pay-offs; see Rabin (1993) or Sugden (2000) for a similar idea. These are pay-offs that are distinguished from the material ones captured in the standard game theoretic representation of an interaction. They arise because players hold beliefs about the fairness of any material outcome which affect their assessment of it.

If fairness provides a motivation in this way, then it becomes important to understand how people come to have ideas regarding what fairness is: what does it consist of and when does it apply? The latter is important because these new theories also typically generate multiple equilibria and so pose the same question regarding equilibrium selection. This paper also reports on an experiment which begins to address this question.

In particular, we amended the HD game by adding a third co-operative strategy.[1] We call the amended game the Hawk–Dove–Co-operate game (HDC). If both players select the 'co-operative' strategy in this new game, the outcome is symmetrical and Pareto-dominates all three of the game's Nash equilibria. However, the co-operative strategy is not part of any equilibrium according to

[1] Note that this third strategy was never related to subjects as 'co-operative'. Strategies were only referred to by their number.

either standard or evolutionary game theory and, from these perspectives, the new game is strategically the same as HD. Mutual co-operation is however a 'fairness equilibrium' in the sense that both Rabin (1993) and Sugden (2000) suggest. We were interested in

(i) whether the mutual co-operation outcome persisted in repeated play and
(ii) whether this fairness equilibrium was still selected if, *outside the HDC game*, players have experienced a convention which gives one of them an arbitrary advantage.

The thought here is that if Hume's ideas are right, then the ideas of fairness associated with an asymmetric convention in the play of HD will militate against the symmetric fairness equilibrium of mutual co-operation when HDC is played.

Thus the paper makes two contributions to the literature. It reports on an experiment that is designed to test

(*a*) for the emergence of a convention based on arbitrary colour assignments which enables equilibrium selection in the HD game and
(*b*) for the endogenous generation of normative expectations in HD which affect play in HDC.

The former addresses a prediction in evolutionary game theory and the latter addresses some particular concerns in behavioural economics with respect to the formation and influence of 'psychological' pay-offs. The organisation of the paper is as follows. Section 1 sets out and considers the two games in more detail. Section 2 describes the experiment. Section 3 gives the results, Section 4 offers an interpretation and section 5 concludes.

1. The Hawk-Dove Game and an Amended Version

Table 1 gives an HD game. There are two common analyses of this game when it is played repeatedly and anonymously: the standard (or conventional) approach and an evolutionary version.

Standard game theory assumes fully rational agents and finds that HD has three Nash equilibria, two in pure strategies (h,d) and (d,h) and one in mixed strategies ($p = 1/3$, where p is the probability of an 'h' choice).[2] The evolutionary approach, on the other hand, assumes non-rational players who gravitate toward the strategy with the highest pay-offs. In the biological interpretation of an evolutionary process, the gravitation occurs because high pay-offs confer reproductive success; whereas, in the social interpretation of the process, it happens because people learn from the success of others. We consider two possible types of evolution:

(i) *One-dimensional evolution* applies to an homogeneous population. Since all members are identical in every way, the evolution of strategies is the same for all members.

[2] Since individual behaviour is unobservable, and there is no room for trigger strategies to develop due to replacement of one's opponents after each round, players cannot invest in some reputation. Thus, each round resembles a one-shot game.

Table 1

The Hawk-Dove Game

	h	*d*
h	−2, −2	2, 0
d	0, 2	1, 1

(ii) *Two-dimensional evolution*: All members are identical, with one small exception. Some have one arbitrary feature, the remainder the other. This difference, though arbitrary, endows the evolutionary process with a second dimension because the fact that each player possesses one of two distinguishing (and observable) features makes it possible for individual behaviour to be conditioned on one's own feature (as well as on the feature of one's opponent). The result is that the strategy which gathers popularity among members of one group may be different from that which is established in the other.

Under one-dimensional evolution, there is a unique evolutionary equilibrium: *the proportion or probability* (p) *of players choosing 'h' equals* $1/3$. This follows because the average return to a person playing 'h' will be greater (less) than playing 'd' for any value of $p < 1/3$ ($p > 1/3$). Consequently more (less) players will opt for 'h' if $p < 1/3$ ($p > 1/3$) and p will rise (fall). Therefore, p will only be stable when it equals $1/3$, a value which coincides with the Nash equilibrium in mixed strategies.

With two-dimensional evolution, there are two evolutionary equilibria. Suppose the population is divided into two equally sized groups by an arbitrary colour identification: members are somehow labelled either blue or red. In meetings between players of different colour, the two evolutionary equilibria are: *'red plays h and blue plays d'* or *'red plays d and blue plays h'* (Weibull, 1995; Friedman (1996).[3] The key to this result is that strategies can be conditioned on colour in cross-colour meetings. Suppose that, at the outset and for no particular reason, the frequency of 'h'-play by blue people falls below $1/3$ (and happens to be less than the frequency of 'h'-play by the reds). Then red persons will discover that, when matched against a blue person, the return to 'h' exceeds that of 'd' and thus 'h'-play among red people will increase. This will reinforce the relative attractiveness of 'd'-play for blue people in cross-colour encounters. In the end, all blue players will be playing 'd' and all red players 'h'.[4] Meanwhile the unique evolutionary equilibrium for meetings between players of the same colour coincides with the one-dimensional equilibrium ($p = 1/3$).

The evolutionary equilbria in mixed colour meetings that result in (h,d) or (d,h) can be interpreted as conventions (Lewis, 1969). Indeed they constitute a form of

[3] Our choice of colours is not random. Mehta *et al.* (1994) report on a laboratory experiment of the 'name any colour' type which shows that blue and red are, roughly, equally salient. This is important because we wanted to preclude an additional source of salience; eg a situation in which *at the very outset* players of one colour (ie the one with higher salience) are seen as more likely to play aggressively as those of the other (ie the less salient) colour.

[4] The opposite of course would be true if, at the outset, the frequency of 'h' among the 'reds' were to fall below both $1/3$ and that of the 'blues'.

discriminatory convention in the sense that they assign each person, on the basis of his or her colour, to either the hawkish or dove-like role[5] and this results in people of one colour enjoying much higher pay-offs than those of the other for reasons which have nothing to do with superior rationality, information or contribution.

One objective of the experiment is to test for whether a discriminatory convention of this sort develops when each player is identified by an arbitrary blue or red colour. We call this the *discrimination* hypothesis. The null hypothesis, supported by standard game theory and one-dimensional evolution, is that colour labels will *not* influence behaviour. The alternative hypothesis, supported by two-dimensional evolution, is that players will, eventually, make use of the extraneous information of colour labels to build a discriminatory convention.[6]

The second game (HDC) in the experiment is set out in Table 2. The original HD game has been amended by the addition of a third 'co-operative strategy', '*c*', for each player. This third strategy is not part of any equilibrium: it will not be played in a repeated setting according to standard game theory and will disappear in the evolutionary version.

Nevertheless, there is some experimental evidence – see Camerer and Thaler (1995) for a survey – suggesting that strategies similar to '*c*' survive (eg the co-operative strategy in the prisoner's dilemma).

One explanation of the persistence of co-operative play in interactions like HDC turns on the identification of 'psychological' pay-offs that come from the symbolic properties of an outcome (its 'fairness', 'goodness' etc). For example, Rabin (1993) assumes agents who derive utility not only from expected monetary returns but also from a perception that they acted fairly. In his account, the perception of fairness (and hence the psychological pay-off) depends on reciprocating 'kindness' (or 'unkindness'). To make such judgements, each player needs to form second-order beliefs regarding what his or her opponent expects him or her to play. So, for instance, suppose Cressida is playing HDC against Troilus and contemplates playing '*c*' because she predicts Troilus will also play '*c*'. Her utility pay-off from outcome (*c*,*c*) varies depending on what she thinks about Troilus's *motivation* for playing '*c*'. 'Is Troilus about to play '*c*' by accident? Or is he also expecting me to play '*c*'?' In the latter case, Troilus's choice of '*c*' contains a measure of kindness to Cressida: given his second-order beliefs that Cressida was going to play '*c*', he *could* have collected pay-off 4 (by playing '*h*') but settled for pay-off 3 and this enables Cressida to enjoy 3

Table 2

The Hawk-Dove-Co-operate Game

	h	*d*	*c*
h	−2, −2	2, 0	4, −1
d	0, 2	1, 1	0, 0
c	−1, 4	0, 0	3, 3

[5] 'The intuition is that a stable mixture of hawks and doves will evolve in a single population, but with two interacting populations, one will become all hawks and the other all doves' (Friedman, 1996, p. 7).

[6] It is worth remarking that there are other possible explanations for the emergence of such a convention. For instance, it might be explained by a version of variable frame theory (Bacharach and Bernasconi, 1997).

rather than –1. In analogous manner, when she plays '*c*', expecting Troilus to play '*c*', she also shows kindness to Troilus. When kindness is reciprocated in this way, Rabin argues that Troilus and Cressida both enjoy a 'psychological' pay-off and when these pay-offs are suitably weighted with the material ones, it is possible for (*c*,*c*) to become what is called a 'fairness' equilibrium.[7]

Appendix F offers a full account of Rabin's theory and Table 3 (page 687) contains a summary of its predictions for HD and HDC.[8] The point to note here is that Rabin's theory depends on his definition of 'kindness' shown by Troilus to Cressida and *vice versa*. Rabin assumes that Troilus's perceived kindness depends on a comparison of Cressida's actual pay-offs from a strategy relative to some assumed reference point. This reference point is given exogenously and defines, in effect, an entitlement for Cressida.[9] When Troilus enables Cressida to obtain something more than this entitlement, he is being 'kind'. Our suspicion is that when people are motivated by such 'psychological' pay-offs, perceptions of entitlement may be formed in a more complex manner than this; and this is why we have included this game in the experiment.

In particular, we are interested in an argument from Sugden (1986) which suggests that ideas regarding what is 'fair' or 'just' may evolve endogenously in the course of social interaction. Sugden follows Hume (1740) by suggesting that, when a convention emerges in a game like HD, it can induce a set of supporting normative ideas: that is, ideas that make the arrangement seem 'just' or 'fair' or some such. It is as if people find it difficult to accept that the convention is, in some sense, arbitrary while also being discriminatory. '*Red plays hawk and blue plays dove*' would perform just as well as a convention as '*blue plays hawk and red plays dove*'. However, the selection of one of these conventions makes a big difference to who receives the most benefit and this seems to cause dissonance. So people remove the dissonance by finding, or inventing, additional principles that will justify the actual convention because it is 'just', 'fair' or some such. If this is the case, then it seems that play of the HD game may induce different ideas regarding entitlements to the play of the HDC game. This is because a convention in HD is inherently discriminatory while it seems from earlier experiments that people are attracted (possibly on grounds of fairness) to the symmetric (*c*,*c*) outcome in games like HDC.

Such a tension between discriminatory and symmetric ideas regarding what is 'fair' or 'just' could make the play of these games sensitive to the order in which they are played. For example, when HDC is played first, a discriminatory convention is less likely to emerge than when HD is played first. This is because the symmetric ideas which may be encouraged by the presence of the co-operative strategies in

[7] Of course, the darker side of Rabin's (1993) fairness model is that Cressida may also value outcome (*h*,*h*) if she thinks that Troilus played '*h*' not because he anticipated '*d*' from Cressida, but because he expects her to play '*h*' and thus wants to hurt her. Then Cressida may derive more utility from (*h*,*h*) than from (*d*,*h*)! In equilibrium, (*h*,*h*) is sustained by the mutual pleasure of hurting each other.

[8] Sugden (2000) presents a similar argument with respect to how (what he refers to as) *normative expectations* can motivate players and produce a mutual co-operation equilibrium in HDC. However, unlike Rabin (1993), (*h*,*h*) would not be an equilibrium on Sugden's account because his sense of fairness does not spring from a psychological need to 'repay' an opponent for their actions but, rather, from doing what is expected. On the other hand, whereas Rabin rules out (*h*,*d*) as a fairness equilibrium when (*c*,*c*) is one, Sugden does not.

[9] Appendix F outlines Rabin's (1993) definition of each player's different entitlements when she/he chooses '*h*', '*d*' or '*c*'.

HDC could inhibit the growth of the discriminatory convention in the play of HD. Likewise, the symmetric (c,c) outcome is less likely to occur in HDC when it has been preceded by HD as compared with experiments in which subjects played HDC first. This is because the discriminatory ideas that might be encouraged in the play of HD could carry over to the play of HDC and inhibit symmetric co-operation. This is our second hypothesis which we refer to as the *sequence* hypothesis.

To be specific, the null hypothesis here is that the sequence of play of HD and HDC makes no difference to behaviour in either game. The alternative hypothesis is that a discriminatory convention is more likely to emerge when HD is played first and that mutual co-operation will be different when HDC is played second. This is supported by the idea that people are motivated by 'psychological' pay-offs and that the perceptions of entitlements which influence these pay-offs depend both on the presence of extraneous information and can be generated endogenously. The comparison with standard game theory and Rabin (1993) is instructive. Since neither standard game theory nor Rabin's theory has a theory of equilibrium selection to offer us, neither makes a prediction regarding an order effect. So, if there is an order effect, then neither standard game theory nor Rabin (1993) can explain it.

2. The Experiment

Four treatments were used to test the two hypotheses. The subjects played each of the two games (HD and HDC) 32 times under quasi-random matching in all four treatments. The treatments differed in two ways: in terms of

(a) whether or not players were labelled as blue/red, and
(b) whether the 32 rounds of HD preceded, or followed, the 32 rounds of HDC.

In 8 sessions no information about individual opponents was provided. We shall refer to them as the no-colour treatment. In another 24 sessions, the colour treatment, players were assigned a colour label at the beginning of the session and were informed of the colour label (blue or red) of their opponent. It is by observing behavioural differences between the colour and no-colour treatments that we test the *discrimination* hypothesis.

In 16 of the 24 Colour sessions, the 32 rounds of HD preceded the 32 rounds of HDC (the *HD-HDC-Clr* treatment). In the remaining 8, the order of play was reversed (the *HDC-HD-Clr* treatment). Similarly in 4 of the 8 no-colour sessions HD preceded HDC (the *HD-HDC-NClr* treatment) while in the remaining 4 no-colour sessions HDC was played first (the *HDC-HD-NClr* treatment). Appendix A offers full details. It is by observing differences in the pattern of play between the *HD-HDC-Clr* and the *HDC-HD-Clr* treatments that we test the *sequence* hypothesis.

2.1. *The Experimental Design*

The 640 subjects came mostly from the student population at the University of Sydney over a period of two years. The group size in each of the sessions varied from

16 to 26 (see Appendix A for details). Once seated in front of their terminal, they were asked to consult on-screen instructions[10] and to ask questions of clarification.

The instructions informed players of the following: the total number of rounds (64); the pay-off matrix of the first game (either HD or HDC); that the game would be amended after 32 rounds to another game (without telling them what the emendation would be) which would also be played 32 times; that, at the end of the session, each player would collect in Australian dollars the sum of her or his numerical payoffs from each round;[11] that one player would win an additional A$10 from a lottery at the end of the session in which his/her chances would be proportional to how many correct predictions of his/her opponents' choice he/she made; that in each round they would be drawn at random against any player in the group (regardless of colour in the 'Colour' treatments) *except that they would never be drawn against the same player twice in a row.*[12]

Following a dry run of four rounds of the first game,[13] the session-proper commenced. In the colour treatments, the colour labels were distributed just before the dry run took place. (Note that the on-screen instructions made no mention of colour labels.) An instructor in full view of players showed them a pack of cards equal in number to that of players. One side of each card was white and the other was either blue or red (half of the cards were blue and half were red). To guarantee that the randomness of the colour distribution was common knowledge, the pack of cards was shuffled in public view. Then the instructor walked over to each subject inviting him or her to pick one at random (before choosing a card subjects could only see the white side of the cards on offer). Once they had collected their coloured card, their screen requested that they punch in 'b' if their card was blue and 'r' otherwise.

Since the games were symmetric, and to avoid introducing a second discriminant (namely, a row or column) which could have given rise to four-dimensional evolution, in all treatments players were told that they were choosing among the rows.[14] In each round subjects had to make two decisions. The first was to predict the strategy which their opponent would select in that round. The purpose of this

[10] Available on request from the authors.

[11] A minimum payment of A$10 was guaranteed. However, this floor was binding in only 4 out of 640 cases.

[12] As is conventional in the literature, anonymity coupled with random matching and the knowledge that one would never play against the same player twice prevents the game from becoming a repeated game and, instead, renders it evolutionary (in the sense that players on the one hand cannot deploy trigger strategies – which require that the *same* players play repeatedly against one another and strive to build a reputation on eponymity – while, on the other hand, they condition their behaviour to the group's aggregate trends). In fact, the software used a simple algorithm to match players (which of course the players were unaware of). To ensure that in the 'colour' sessions all red players would be matched against a blue player an equal number of times (and *vice versa*), and that the matching protocol would be as close to random (which is what subjects were 'promised' it would be) as possible, the algorithm produced *per player* an equal number of pairings with a player of the same colour as of the opposite one. In aggregate, the algorithm guaranteed that in the 32 rounds of each game (HD and HDC) the distribution of blue-blue, red-red and blue-red pairs would be ¼, ¼ and ½ respectively.

[13] The familiarisation rounds involved the first game of the session (that is, HD in treatments HD-HDC or the HDC game in treatments HDC-HD). Afterwards the computer checked, via two multiple choice questions, whether the players understood the way in which their payoffs would be decided. The session did not begin unless all subjects passed this mini-test.

[14] In a separate set of experiments with a battle-of-the-sexes type of game, we have found that whether a player chooses among the columns or the rows can evolve into a powerful discriminant. For instance, we discovered that in the standard 2×2 version of that game, there was a strong tendency towards the Evolutionary/Nash equilibrium which favours the row players. See Varoufakis (1996).

was to gauge the first order (predictive) beliefs of subjects for later use (see section 4)[15] and, to avoid unmotivated responses, subjects were offered a lottery ticket for every correct prediction.[16] After the predictions of each player were registered, they were then invited to make their own strategic choice.

In the colour treatments, the computer informed players of the colour of their opponent at the beginning of each round. In the no-colour treatments no information was given about one's opponent. When all subjects had registered their predictions (of their current opponent's choice) *and* punched in their choice of strategy, the round was over and their screen would provide the following information:

(i) His/her opponent's choice (and thus his/her pay-off from this round)
(ii) The group's aggregate behaviour in both the last round and for all rounds so far (on average); eg 30% chose '*h*', 60% chose '*d*' and 10% '*c*'
(iii) The running total and the average of *his/her* pay-offs for all rounds so far
(iv) The average pay-offs of the group for all rounds so far.

In colour sessions, players were given additional information:

(v) The aggregate behaviour of all red players and of all blue players separately, both in the last round and for all rounds so far (on average)
(vi) The running average pay-off of blue and of red players separately.

As is common practice in experiments of this type, the purpose of giving feedback to subjects in experiments is to remove sampling error and speed up convergence, thus avoiding the concentration lapses (not to mention spiralling costs) caused by a greater number of rounds. A printout of the screen offering a snapshot of what the players saw during the sessions can be found in Appendix A.

3. Results

The theoretical predictions for our four treatments that come from standard and evolutionary game theory together with Rabin's fairness equilibria are summarised in Table 3.

[15] There are two ways for soliciting expectations about discrete events. One is to ask agents (as we did here) to predict which of the two (three) strategies his/her opponent would choose in HD (HDC). The second way is to invite them to tell us the odds, as they see them. The latter has the advantage of revealing more about the agents' subjective p.d.f. However it suffers from two disadvantages. One is the (usually mistaken) presumption that subjects are familiar with distributions (and that they can express accurately their beliefs in probabilistic terms). The second disadvantage is that, unlike the former technique, it makes it hard to devise a simple reward scheme which will motivate subjects to reveal their expected distribution accurately. In selecting the former we decided to opt for the simplest question (ie which strategy, '*h*', '*d*', or '*c*' do you think is more likely that your opponent will choose?), the simplest payoff-structure (ie if your guess is correct you will increase your chance of winning a prize) and the simplest (to interpret) reply. Since the sample size was large, and the objective was to monitor the *trend* of changes in such predictions (as opposed to their mean and standard deviation), the advantages of discrete predictions were deemed considerable.

[16] The lottery scheme was calibrated in such a way that if one predicted correctly all 64 choices by one's opponents, one would gain a 100% chance of winning A$10 *in addition* to the pay-offs from the games.

Table 3

Predictions of Long Run or Equilibrium Behaviour

The Predictions of Conventional Game Theory

(a) Behaviour will converge on one of the three Nash equilibria available: (h, d), (d, h) or $\Pr(h) = 1/3$.
(b) No prediction regarding order effects.

The Predictions of Evolutionary Game Theory

No-colour treatments	*Colour treatments*
(a) One dimensional evolution will lead to the unique evolutionary equilibrium (also a Nash equilibrium in mixed strategies): $\Pr(h) = 1/3$.	(a) *Different colour meetings*: Two-dimensional evolution leading to a unique evolutionary (pure strategy) equilibrium in which players holding one of the two colours play '*h*' and holders of the other colour play '*d*'
(b) The third strategy '*c*' will fade away in game HDC.	(b) *Same colour meetings*: One dimensional evolution leading to the unique evolutionary equilibrium (also the Nash equilibrium in mixed strategies): $\Pr(h) = 1/3$
(c) No prediction regarding order effects.	(c) Strategy '*c*' will fade away in game HDC. (d) No prediction regarding order effects.

The Predictions of Rabin's Model of Fairness[17]

(a) All cells on the diagonal of the pay-off matrices of games HD and HDC may be observed systematically, in addition to the pure strategy Nash equilibria (h, d) and (d, h).
(b) Outcome (h, h) will occur more frequently in HD than in HDC (or, at least, not less frequently).
(c) Outcome (d, d) will occur more frequently in HDC than in HD (or, at least, not less frequently).
(d) The pure strategy Nash equilibria (h, d) and (d, h) will occur more frequently in HD than in HDC (or, at least, not less frequently).
(e) The use of strategy '*c*' in game HDC will not fade away.
(f) No prediction regarding order effects.

Tables 4 and 5 offer an overview of the experimental data. The data are expressed in percentages rounded-off to one decimal point. The data for the game that was played first are highlighted and the italicised figures signify that the relevant

Table 4

Frequency (%) of Outcomes in all 32 Rounds of Each Game per Treatment

Game Outcomes	*HD*				*HDC*				
Treatment	(h, h)	(h, d)[18]	(d, d)	(h, h)	(h, d)	(d, d)	(c, c)	(h, c)	(d, c)
HD-HDC-NClr	**29**	**39.8**	**31.2**	36.7	9.8	3.7	6	30.2	13.6
HDC-HD-NClr	33	35.6	31.4	**29.3**	**4.3**	**2**	**8.2**	**38.1**	**18.1**
HD-HDC-Clr	***21.4***	***51.8***	**26.8**	*19.2*	*38.7*	2.2	9.3	*20*	10.6
HDC-HD-Clr	26.9	45.2	27.9	**30.1**	**7.1**	**2.1**	**7.2**	**34.7**	**18.8**

[17] The predictions below are derived from Rabin's (1993) model as explained in Appendix F. In brief, if v denotes the marginal importance of money relative to the psychological pay-offs, it transpires that the influence of the psychological pay-offs is a diminishing function of v. For v values below certain thresholds, the diagonal elements of the pay-off matrices become equilibria while the original Nash equilibria (h, d) and (d, h) drop out. Predictions (a)–(e) are based on the implicit hypothesis (consistent with Rabin) that there exists a random (exogenous) distribution of the vs among our subjects. Naturally, as there are multiple equilibria and no theory of equilibrium selection, these predictions are based on the presumption that the likelihood of each equilibrium is proportional to the range of v values which supports it.

[18] Due to the games' symmetry and the fact that all players were choosing among the rows, outcomes off the diagonal are reported as one: e.g. (h, d) data reports on the frequency of both (h, d) and (d, h), etc.

Table 5

Frequency (%) of Strategies in all 32 Rounds of Each Game per Treatment

Game Strategies / Treatment	*HD*		*HDC*		
	'*h*'	'*d*'	'*h*'	'*d*'	'*c*'
HD-HDC-NClr	**48.9**	**51.1**	56.7	15.4	27.9
HDC-HD-NClr	50.8	49.2	**50.5**	**13.2**	**36.3**
HD-HDC-Clr	**47.3**	**52.7**	48.5	*26.9*	24.7
HDC-HD-Clr	48.5	51.5	**51**	**15.1**	**34**

observation in treatment *HD-HDC-Colour* is different from those in the same column (ie of the other treatments) at the 95% confidence level.[19] The data here come from both the early, more 'noisy' rounds as well as the later ones (to which the predictions in Table 3 apply more readily). Nevertheless there are three important results.

- *Result 1*: Treatment *HD-HDC-Colour* stands out in terms of the frequency of the pure strategy Nash equilibrium (*h*,*d*). In both games (HD and HDC), the frequency of the pure strategy Nash equilibrium (*h*,*d*) is significantly larger in this treatment than in the rest. In game HDC, this difference becomes overwhelming.[20]
- *Result 2*: Behaviour in treatment *HDC-HD-Colour* is significantly distinct from that in *HD-HDC-Colour*, and rather similar to that in the no-colour treatments. In particular, the frequency of outcome (*h*,*d*) in *HDC-HD-Colour* is statistically indistinguishable from the two no-colour treatments (and, of course, significantly lower than in *HD-HDC-Colour*).
- *Result 3*: Co-operative behaviour is present in the HDC game in all treatments.

Result 1 is directly relevant to the *Discrimination* hypothesis and is consistent with the two-dimensional evolutionary model.[21] This finding is reinforced by the data in Table 4 showing that the more frequent occurrence of (*h*,*d*) in *HD-HDC-Colour* was achieved, especially in HDC, in spite of the fact that players did *not* play, in aggregate, '*h*' or '*d*' with frequencies significantly different to those in other treatments (see Table 5). It seems, therefore, that there was *something* in *HD-HDC-Colour* that enabled players to co-ordinate their '*h*' and '*d*' choices so as to boost the incidence of outcome

[19] The statistical tests used here need to be qualified. Although common in the experimental literature, they are open to the criticism that they treat as independent what might, after all, be repetitions of a single (or a few) observation(s). [NB this would be indeed true if players converge quickly to a fixed response.] Nonetheless, such criticism is pertinent when the reported statistical significance is marginal. In cases, like ours, where the differences between treatments are large, there is no cause for concern.

[20] In fact, the frequency of (*h*,*d*) in game HDC of *HD-HDC-Clr* is four times greater than the second highest frequency of the remaining treatments. The null hypothesis that the frequency of this pure strategy Nash equilibrium is the same across treatments *HD-HDC-NClr*, *HDC-HD-NClr* and *HDC-HD-Clr* cannot be rejected at the 5% level in either game HD or HDC. By contrast, the null that the frequency of outcome (*h*,*d*) in treatment *HD-HDC-Clr* is the same with that in the other three treatments is rejected for HD at the 5% level and for HDC at the 1% level.

[21] Note that the observations from the no-colour treatments are fully consistent with those reported elsewhere, namely one-dimensional HD play; see, for instance, McDaniel *et al.* (1994).

(h,d) at the expense of (h,h) or (d,d). (Whether this *something* was, in fact, the colour labels is the subject of our convergence analysis below.)

By contrast *Result 2* goes beyond the two-dimensional evolutionary model as it points to a clear order effect. Neither standard nor evolutionary theory can explain why the availability of strategy 'c' from the outset seems to prevent the evolution of discrimination. Some emendation like our *sequence* hypothesis seems necessary.[22]

Likewise *Result 3* is not predicted as an equilibrium outcome by standard or evolutionary game theory, but some care is required here as the play of 'c' could result from errors or in the process of learning adaptively. The result is, however, consistent with Rabin's (1993) hypothesis (see Table 3). We consider these options in more detail in section 4.

To examine whether a discriminatory convention lies indeed behind the greater incidence of (h,d) in the *HD-HDC-Colour* treatment, we use a version of Friedman's (1996) test for convergence. In each session, we compute (separately for HD and HDC) the frequency p that, in cross-colour meetings, *blue plays 'h'* and the frequency q that *red plays 'h'* based on the last 5 rounds. If the null hypothesis that $p = q$ can be rejected, we proceed backwards to identify the round by which the discriminatory pattern observed in the last 5 rounds had settled down. Full details are given in Appendix B, but the idea is to look for the largest number of rounds before the end which would give estimates of p and q which do not differ (at a 95% confidence level) from those values in the last 5 rounds. When $p > q$, then we say blue is advantaged (A) and red is disadvantaged (D) and *vice versa*.

The table in Appendix B gives the results for whether convergence occurred in each of the sessions, which colour was advantaged by it, and by which round convergence was achieved (if it was). In 15 of the 16 sessions of treatment, *HD-HDC-Clr* convergence occurred within, on average, 15.9 (out of 32) rounds of HD. By contrast, only one of the 8 sessions in the *HDC-HD-Clr* treatment showed convergence in the first part, the HDC game. Hence, there is evidence from the study of convergence which points to the emergence of a discriminatory convention in *HD-HDC-Clr* treatment; and there is evidence that availability of 'c' from the outset prevented the evolution of a similar pattern in HDC in 7 out of the 8 *HDC-HD-Clr* sessions. In short, the sequence of play does appear to matter.

Table 6 summarises this evidence on the two hypotheses from data based on observations from all 32 rounds of each game. It shows that

(a) the null hypothesis based on standard game theory (ie that behaviour in *HD-HDC-Clr* is indistinguishable from *HD-HDC-NClr*) is rejected
(b) the null hypothesis (i.e. that the sequence of play of HD and HDC makes no difference to behaviour in either game) is also rejected.

Instead, there is evidence that is consistent with the emergence of a discriminatory convention based on colour identification in the colour treatment and evidence

[22] Perhaps the availability of strategy 'c' does not derail the evolution of discrimination but, instead, slows it down. Indeed, it is possible to show in the context of an evolutionary analysis of HDC that there exist trajectories which, initially, take the evolutionary process away from the equilibrium (eg by boosting the frequency of co-operative play) before returning to it.

Table 6

Testing the Hypotheses on Aggregate Data[23]

	Null hypothesis; Alternative hypothesis in parenthesis	Sample sizes	p-values HD	p-values HDC
Discrimination Hypothesis[24]	In *HD-HDC-Clr*, Freq of '*h*' by blue players = (≠) Freq of '*h*' by red players	10,560 choices by blue and 10,560 by red players	0.04*	0.008*
	Freq of '*h*' in *HD-HDC-Clr* = (≠) Freq of '*h*' in HD-HDC-NClr	10,560 choices in *HD-HDC-Clr* and 2,816 in *HD-HDC-NClr*	0.02*	0.009**
Sequence Hypothesis	The proportion of sessions in which *Discrimination* evolved is (is not) identical across *HD-HDC-Clr* and *HDC-HD-Clr*	16 sessions of *HD-HDC-Clr* and 8 of *HDC-HD-No Clr*; discrimination was observed in 15 of the former and 1 of the latter	0.002**	

that the presence of '*c*' at the outset inhibits the emergence of a discriminatory convention.

In the remainder of the paper, we focus on the 16 sessions where we have evidence that a discriminatory convention emerged well before the half point of the session (and regardless of which game was played first). Of these 16, 15 were sessions of the *HD-HDC-Clr* treatment and only one of the *HDC-HD-Clr* treatment (see Appendix B for details). Table 7 compiles data from these 16 sessions *from the last 11 rounds of HDC only*; that is, the reported frequency of outcomes emerged well after the convention had begun to take hold. Since the discriminatory conventions were well established by the time the Table 7 dataset was compiled, we could identify whether each player was either advantaged (*A*) or disadvantaged (*D*) by the convention and so we plot the frequency of outcomes depending on whether the meeting is between mutually advantaged (*A*) or disadvantaged (*D*) players or between an advantaged (*A*) and a disadvantaged (*D*) player.

Table 7 reveals again the influence of convention. In colour sessions in which a convention did *not* become established[25] (see last row of Table 7), the pure strategy Nash equilibrium (*h*,*d*) occurs only 5.6% of the time. In sharp contrast, in the sessions where a convention did emerge, we find that when *A*-players met

[23] *p-values* The reported *p*-values refer to the empirical probability that the value of the relevant test statistic is as extreme or more extreme than its observed value assuming the null hypothesis to be true. For example, the p-value of 0.002 reported for the *sequence* hypothesis means that the null of order-independence in the colour sessions can be rejected with 99.8% confidence.

Test statistics Two pooled *t*-test statistics were used in connection to the *discrimination* hypothesis. One tested whether the frequencies with which the blue and the red players chose strategy '*h*' in *each session* of the *HD-HDC-Clr* treatment were equal; see the p-value marked with (*). The other compared the frequencies of '*h*' in *HD-HDC-Clr* with that in *HD-HDC-NClr*; the relevant p-value is marked with (**). The p-value, namely the *sequence* hypothesis is based on a simple two-sample pooled *t*-test.

[24] Note that this test of our *discrimination* hypothesis is considerably biased in favour of the null hypothesis: The data used contain not only the early rounds (during which a fledgling convention had no time to emerge) but also the same-colour meetings in which the *discrimination* hypothesis does not predict differences in '*h*'-play between the red and the blue players. And yet despite of all this 'noise' which ought to have made it harder to reject the null, in treatment *HD-HDC-Clr* (see Table 6), the null was rejected handsomely.

[25] Largely because of the availability of '*c*' from the outset.

Table 7

Data from the Last 11 Rounds of Game HDC

Mean outcome frequencies in the last 11 rounds of HDC in the 16 Colour sessions in which players of one colour (A) gained an advantage at the expense of players of the other colour (*D*)	(h, h)	(h, d)	(d, d)	(c, c)
Meetings between two A players	*51.2%*	8.9%	1.81%	4%
Meetings between two D players	2.1%	3.5%	0.5%	*89.9%*
Meetings between an A and a D player	8.2%	*81%*	0.4%	0.5%
Comparable frequencies in the 8 colour sessions in which *no* discriminatory convention was established	22.1%	5.6%	3.1%	8.2%

Note: Italicised frequencies exceed the other frequencies in the same column with at least 99% probability.

D-players, the pure Nash equilibrium of (h,d) is achieved with a very high frequency (81%).

Table 7 also reveals another interesting difference. We find that in those sessions where the discriminatory convention emerged, there is a conspicuously high incidence (almost 90%) of the co-operative (c,c) outcome between *D*-players. In comparison, there is mutual co-operation between *A*-players only 4% of the time and there is a negligible amount of co-operation between *A* and *D*-players. Likewise, when no convention emerges the level of mutual co-operation is strikingly lower at 8.2%.[26] In other words, it seems that the part of the *sequence* hypothesis relating to co-operation receives support from the data in the sense that when a discriminatory convention emerges, it is associated with very high levels of co-operation between the *D*-players. We focus on this result in the next section.

The combined influence of the discriminatory convention and this sequence effect can be seen from another angle in Table 8. This gives an analysis of the distribution of average pay-offs. It shows that, over all rounds of treatment *HD-HDC-Clr*, A-players received 90% of their money from meetings with *D*-players. On the other hand, 71.8% of *D*-players' winnings came out of meetings with other *D*-players. Put differently, whereas only 5.8% of *A*-players' earnings were due to co-operation with other *A*-players, *D*-players received 61.7% of their total pay-out from co-operating with one another.

[26] The hypothesis that *D*-players are more co-operative than *A*-players is even supported by the aggregate, noisy data (ie data from all 32 rounds of HDC) as this table demonstrates:

Null hypotheses	p-values; sample sizes in brackets
$\mathrm{Fr}(c\|[A,A]) = (<)\mathrm{Fr}(c\|[D,D])$	0.04♦ (4,928)
$\mathrm{Fr}(c\|[D,D]) \to a$ *and* $\mathrm{Fr}(c\|[A,A]) \to b$ where $a = (>)0$ and $b = (>)0$	0.000♦♦ (4,928)

$\mathrm{Fr}(d|[A,A])$ is the frequency with which strategy '*c*' was chosen in meetings between two *A*-players, etc. The p-values are underpinned by a similar pooled *t*-statistic which tests the null that the frequency of strategy *c* is the same in *A*-player meetings compared to *D*-player meetings (the relevant p-value is marked by ♦) *and* that the frequency of successful co-operation among *D*-players or among *A*-players vanishes (the relevant p-value is marked by ♦♦). (Note that a Wilcoxon non-parametric test, not reported here, gave similar results.)

Table 8

Average Payoffs per Round (in Australian cents) of A-players and D-players in all 32 rounds of HD and HDC in treatment HD-HDC-Colour

	Game HD		Game HDC	
	A-players	*D*-players	*A*-players	*D*-players
Meetings between an *A* and a *D* player	66.3	21.7	137.8	39.7
Meetings between two *A* players	7.3	–	16.2	–
Meetings between two *D* players	–	19.6	–	101.3
Average	*36.8*	*20.7*	*77*	*70.5*

4. Why Did Co-operation Occur Among the 'Disadvantaged'?

From the perspective of standard game theory, there seem to be two possible ways of explaining the high incidence of co-operation among *D*-players. One is to appeal to the heightened kind of rationality which can sustain co-operation among a sub-group by some version of punishment (or trigger) strategies. The difficulty with this interpretation is twofold. First, there remains a question regarding why it is only the sub-group of *D*-players who manage to achieve co-operation in this way. Secondly, under any version of a punishment (or trigger strategy) when the game has a finite horizon, players should abandon co-operation in the last round of the HDC game. However, the null hypothesis that the frequency of '*c*' play by *D*-players in the last round remains the same as that in the previous 31 rounds cannot be rejected at the 5% level in favour of the alternative hypothesis that it fell (in fact it rose slightly).

The other is to appeal to some kind of bounded or adapted rationality. Suppose for instance, there is inertia with respect to strategy selection such that once an *A*-player learns to play '*h*' and the *D*-player learns to play '*d*' in cross-colour encounters, they unthinkingly do the same in same-colour matches. This would explain the high incidence of (*h*,*h*) among *A*-players but it would not explain why (*c*,*c*) results among *D*-players. Perhaps the *D*-players block the '*h*' strategy in mutual encounters (since they do not use it) so that they see a 2×2 version of the HDC which is a pure co-ordination game; see Bacharach (1997) for a variable frame model of co-operation. In this co-ordination game, (*c*,*c*) could become focal on the basis of Pareto and risk dominance and, once established, it just becomes the habit of *D*-players to play '*c*' with each other. The difficulty with this type of argument is that it presumes 'adaptive' players *unthinkingly* use particular strategies once they have been assigned to either the 'advantaged' or 'disadvantaged' role and the data casts some doubt on this.[27]

[27] Notice that such inertia is irrational. Instrumentally rational players (ie those capable of maximising their own pay-offs given their information) would follow an emerging convention only in cross-colour matches. Why? Because in the absence of any guarantees of consistently aligned beliefs, the discriminatory convention offers them useful information about their opponent's likely beliefs and actions. However, in same-colour matches, they are useless. Therefore, only by mistake will pay-off maximisers allow habits which took shape in cross-colour meetings to spread into same-colour ones. Such inertia, or reinforcement, presumes that players pay no attention to the outcomes of strategies that they did not choose. For an interesting discussion, see Erev and Roth (1998) and Erev *et al.* (1999).

Table 9 is drawn from the last one-third run (11 rounds) of treatment *HD-HDC-Clr* and gives the prediction-choice combination for both *A*-players and *D*-players. The first row reports that *D*-players predicted their opponent would choose '*c*' 879 times. In 861 out of those cases, they chose '*c*' themselves. *A*-players predicted '*c*' 789 times, but only responded with '*c*' in 31 cases (see row 3). In meetings with opponents bearing the same colour label as themselves, *D*-players co-operated almost every time they had predicted '*c*' (ie with frequency 98.7%, see row 5). When they had not predicted '*c*' by a fellow *D*-player, they played '*c*' 43.3% of the time (row 6). The latter is a high figure which provides some succour for the habit hypothesis, but since it is under half the figure for when they expected their fellow *D*-player to choose '*c*', it seems that something more than a thoughtless attraction to '*c*' explains behaviour here. Likewise, although *A*-players in their mutual meetings are not attracted very often to play '*c*', its frequency is higher when an *A*-player expects the other *A*-player to choose '*c*' (24.7%, see row 7) compared with when they do not expect '*c*' (9%, see row 8).

Likewise, rows 9–12 caution against this adaptive explanation. *D*-players seem to have thought quite carefully before attempting to co-operate. When they played against *A*-players whom they thought would *not* co-operate, they only

Table 9

The Prediction-Choice Combinations of Subjects in the Last One-third-run (11 rounds) of HDC in Treatment HD-HDC-Colour

	Player's colour	Opponent's colour	Player predicted opponent would play strategy	*AND* then played strategy	Conditional freq (*)	%	p-values < 0.001 (**)
1	*D*	Any	*c*	*c*	861/879	98	
2	*D*	Any	~*c*	*c*	31/789	3.9	
3	*A*	Any	*c*	*c*	59/1,174	5	
4	*A*	Any	~*c*	*c*	82/2,603	3.2	
5	*D*	*D*	*c*	*c*	830/841	98.7	
6	*D*	*D*	~*c*	*c*	29/67	43.3	
7	*A*	*A*	*c*	*c*	20/81	24.7	
8	*A*	*A*	~*c*	*c*	74/826	9	
9	*D*	*A*	*c*	*c*	31/38	81.6	
10	*A*	*D*	*c*	*c*	39/1,093	3.6	
11	*D*	*A*	~*c*	*c*	2/722	0.2	
12	*A*	*D*	~*c*	*c*	8/1,777	0.5	
13	*D*	*D*	*h*	*h*	53/67	79	
14	*A*	*A*	*h*	*h*	602/826	72.9	
15	*A*	*D*	*h*	*h*	23/140	16.4	
16	*D*	*A*	*h*	*h*	28/1,762	1.6	

(*) This column refers to the frequency of particular *combinations* of expectations and choices. For example, the first row reports that, in the last 11 rounds of HDC, there were 879 occasions when *D*-players predicted that their opponent would play '*c*'. Of those 879 instances, *D*-players decided to respond to that prediction by playing '*c*' 861 times. The sixth row reports that there were 67 occasions when, in a meeting between two *D*-players, a *D*-player did not predict '*c*' but played '*c*' regardless 29 (out of those 67) times.

(**) The p-values indicated here by the arrows relate to the null that the two frequencies linked by the arrows are equal.

chose '*c*' in 2 out of 722 cases (row 11); whereas, when they expected that the *A*-player would choose '*c*', they co-operated in 31 out of 38 cases (row 9). Again, this hardly accords with the view that *D*-players were thoughtlessly locked into playing '*c*'. Turning to *A*-players, their propensity to co-operate with a *D*-player was also influenced distinctly by whether they expected '*c*' or not (rows 10 and 12).

Since standard game theory does not seem able to provide convincing explanations of the persistence of co-operation (especially among *D*-players), we now turn to explanations which postulate psychological pay-offs. The earlier discussion (Section 2) indicated how the Rabin model can explain co-operative behaviour in HDC and the conflict outcome (*h*,*h*). Its drawback is that it cannot account for differences in the frequency of co-operative moves between our *A*-players and *D*-players. To make this possible, we would have to amend Rabin's model so as to explain why (*c*,*c*) is selected as a fairness equilibrium among *D*-players but not among *A*-players.

One way of achieving this would be to assume that, while playing HDC in treatment *HD-HDC-Clr*, agents' normative beliefs on entitlement reflect not just the structure of the pay-off matrix – as Rabin (1993) – assumes but, additionally, their role in the discriminatory convention which emerged in the earlier play of the HD game. So, *A*-players might have higher normative expectations regarding entitlements than *D*-players following the play of HD (see the average pay-offs for the HD part of the game reported in Table 8). If this was the case, then (*c*,*c*) could be a 'fairness' equilibrium for *D*-players but not for *A*-players. Instead *A*-players with higher normative expectations may find themselves locked into a nasty (unkind) fairness equilibrium with players of the same colour. In such an equilibrium, they anticipate that their *A*-opponents are about to harm them by playing '*h*' and, to avoid the unfairness of repaying nastiness with kindness (or even with normatively neutral behaviour), they respond to a probable '*h*' with an '*h*'.

As suggested earlier, this kind of endogenous generation of entitlements follows a line of argument in Sugden (1986). It is not implausible given what is known from other experiments (Babcock *et al.*, 1995; Asdigan *et al.*, 1994; Schotter *et al.*, 1996; Binmore and Samuelson, 1993)[28] and it is a natural extension in some

[28] It is not unusual for players belonging to different groups to entertain different perceptions of fairness. For instance, commenting on the data from their dispute-resolution experiment, Babcock *et al.* (1995) conclude thus: 'Even when the parties have the same information they will come to different conclusions about what a fair settlement would be and base their predictions of judicial behaviour on their own views of what is fair.' Asdigan *et al.* (1994) report the well-known fact that men and women rationalise by means of different principles of distributive justice their different socio-economic status as well as that of others. See also Kahn *et al.* (1980), Major and Adams (1983) and Major *et al.* (1989). Schotter *et al.* (1996) suggest, in effect, that such ideas of fairness may be endogenously generated. In an ultimatum game experiment involving 8 pairs of players, the 4 proposers who gained most money (out of the 8 proposers in each session) were given the opportunity to play again (against another responder). In these sessions, the responders (who knew that the proposers were competing against each other) accepted, on average, lower offers than in sessions where the proposers did not compete. Thus, it seems that players are prepared to accept a lesser position if there is some rationale for it. Likewise, Binmore and Samuelson (1993) report that, in the context of ultimatum games, the normative expectations of responders and proposers change at different speeds due to the fact that the former have less to lose from rejecting unfair offers by the latter.

respects of what evolutionary theory suggests regarding the evolution of positive (ie predictive) beliefs into normative beliefs concerning entitlements. Our evidence seems to be adding to this line of thinking.[29]

Nevertheless, the argument is, at best, suggestive. There are tricky issues of detail concerning precisely how entitlement norms evolve which need to be addressed. Furthermore, an appeal to the motivational force of an evolving set of psychological pay-offs is not the only possible way to account for co-operative behaviour among the *D*-players. For instance, it might be possible to argue that *D*-players 'group identify' and so adopt a form of team reasoning which produces co-operation; Bacharach (1999) might explain this as a result of the 'common fate' hypothesis of group identity formation. The point of the argument in this section, then, is simply to lay the ground for a more thorough investigation along these lines because it seems that standard game theory cannot explain the co-operative behaviour among *D*-players while some kind of evolving fairness equilibrium or evolving group identification process could.

5. Conclusion

This paper reports on an experiment with two striking patterns of behaviour: the quick emergence of a relation of dominance in a repeated Hawk–Dove game associated with purely conventional labels; and a tendency for the subjects with subservient labels to co-operate with each other. The first of these bears out the predictions of evolutionary game theory. The second cannot be explained by either standard or evolutionary game theory or Rabin's psychological theory. One possible explanation, however, comes from an amended version of Rabin's (1993) model. If the convention of dominance establishes a norm of different entitlements for those with different labels, then this norm could define a 'fairness' equilibrium among those with a subservient label which involves mutual co-operation. With this interpretation of the matter, the paper not only finds evidence to support an influence from 'psychological pay-offs' but also that they are affected by the presence of a discriminatory convention. This is an important result, not least because it throws new light on Aristotle's famous maxim with which we began the paper.

However, while there is some evidence that supports this interpretation, it is not the only possible one; and there are other aspects of behaviour in the experiment which need explanation. For example, there is the result that plots the reverse influence (how the initial presence of an 'irrelevant' co-operative strategy can inhibit the emergence of a discriminatory convention in the HDC version of the game). It is possible that this too may be capable of explanation through a related amendment to Rabin's original hypothesis, but, in the absence of a more general

[29] Our data on subjects' point estimates of their opponent's choice, though not presented here due to space restrictions, show unequivocally that, as convergence to the discriminatory convention was approaching, our players predicted the observed behavioural patterns rather accurately. For example, *D*-players (*A*-players) increasingly predicted a higher (lower) frequency of *h* if their opponent was of the opposite colour. *D*-players anticipated a higher (lower) degree of co-operativeness from opponents of the same colour than *A*-players.

theory of entitlements, this is not clear. In short, something interesting is going on, Aristotle's maxim is suggestive and more work needs to be done.

University of East Anglia
University of Sydney and University of Athens

Date of receipt of first submission: November 1998
Date of receipt of final typescript: July 2001

Appendix A. The 32 sessions of the 4 treatments

Abbreviations of the four treatments:

Treatment	1st Game (32 rounds)	2nd Game (32 rounds)	Colour labels assigned?
HD-HDC-NClr	HD	HDC	No
HDC-HD-NClr	HDC	HD	No
HD-HDC-Clr	HD	HDC	Yes
HDC-HD-Clr	HDC	HD	Yes

In each treatment subjects played the 1st game 32 times and then played the 2nd game another 32 times. Below the sessions are listed in chronological order. Column *N* denotes the number of subjects in each session.

	Treatment	*N*		Treatment	*N*		Treatment	*N*
1	HD-HDC-NClr	24	12	HD-HDC-Clr	18	23	HD-HDC-Clr	18
2	HDC-HD-NClr	16	13	HDC-HD-Clr	18	24	HD-HDC-Clr	16
3	HDC-HD-NClr	22	14	HDC-HD-Clr	20	25	HD-HDC-Clr	22
4	HDC-HD-NClr	22	15	HD-HDC-Clr	18	26	HDC-HD-Clr	18
5	HD-HDC-NClr	18	16	HD-HDC-Clr	16	27	HD-HDC-Clr	22
6	HD-HDC-NClr	24	17	HDC-HD-Clr	20	28	HD-HDC-Clr	16
7	HD-HDC-NClr	22	18	HDC-HD-Clr	24	29	HD-HDC-Clr	26
8	HDC-HD-NClr	16	19	HD-HDC-Clr	24	30	HD-HDC-Clr	18
9	HDC-HD-Clr	18	20	HD-HDC-Clr	16	31	HD-HDC-Clr	22
10	HD-HDC-Clr	16	21	HD-HDC-Clr	20	32	HD-HDC-Clr	26
11	HD-HDC-Clr	26	22	HD-HDC-Clr	16			

Treatment	No. of sessions	No. of players	Interactions per game
HD-HDC-NClr	4	88	1,408
HDC-HD-NClr	4	76	1,216
HD-HDC-Clr	16	330	5,280
HDC-HD-Clr	8	146	2,336
Total	*32*	*640*	*10,240*

An example of the screen that subjects faced in the fourth round of HD-HDC-Colour is shown below.

The Game - Round 4 of 32

FOR THIS ROUND YOU HAVE BEEN MATCHED (RANDOMLY) WITH A RED PLAYER

		THE OTHER PLAYER	
		1	2
YOU	1	-$2,-$2	$2,0
	2	0,$2	$1,$1

Your payoffs so far: $3
Your average payoffs so far: $1

Information

Frequency of previous choices	**Strategies 1**	**2**
By the whole group (including yourself) in the LAST round	34%	66%
By the whole group (including yourself) in ALL 3 previous rounds	39%	61%
By the blue players (including yourself) in the LAST round	43%	57%
By the blue players (including yourself) in ALL 3 previous rounds	40%	60%
By the red players in the LAST round	25%	75%
By the red players in ALL 3 previous rounds	38%	62%

PLEASE: **Predict the choice that the player you have just been randomly matched with will make in this round.** [Recall that if you predict correctly, you will win, in addition to your money pay-offs from this round, a lottery ticket. At the end of the session, $10 will be given to the player with the lucky ticket. The more lottery tickets you collect the greater the chances of winning the $10.]
Punch in number 1 if you think that she/he will choose strategy 1, or 2 if you think that she will choose strategy 2.

Once the player made his/her prediction, the final paragraph disappeared from the screen and the following emerged:

NOW CHOOSE YOUR OWN STRATEGY: **Punch in number 1 if you wish to select strategy 1, or 2 if you prefer strategy 2.**

Appendix B: The emergence of 'advantaged' and 'disadvantaged' colours in the colour treatments (Sessions 9 to 32)

Description of the algorithm used to establish whether (and if so in which round) discrimination emerged

Let p = Freq(*Blue*→h) and q = Freq(*Red*→h) denote the frequency of event '*Blue* (or *Red*) player chose h in some round of a game'.

STEP 1: In each session, compute (separately for HD and HDC) p and q from the last 5 rounds of the game. If $p > q$, set A = *Blue* and D = *Red* or *vice versa*. Let $\pi = \text{Freq}(A \rightarrow h)$ and $\theta = \text{Freq}(D \rightarrow h)$. If the null that $\pi = \theta$ can be rejected with 95% confidence (in favour of the alternative hypothesis that $\pi > \theta$), then STOP. (For if it cannot be rejected, then no convergence was achieved by the end of the game's 32 rounds.) If it can, proceed to STEP 2 to identify the round by which the discriminatory pattern which was observed over the last 5 rounds had settled down.

STEP 2: Following Friedman (1996), this convergence criterion was used: $(1/L)\sum \text{SupNorm}\{\pi' - \pi, \vartheta' - \vartheta\} \leq \epsilon$ where L is the length of run R under scrutiny. Values π and θ, as before, were computed over the last 5 rounds of the game in question. Values π' and θ' were computed over the run of length L. At first we set L = 6 and chose as our 6 observations the last 6 rounds of the game. Thus run R initially included the last 6 rounds of each game in each session. If the criterion was

met for the chosen value of ε (see below for an explanation of how ε was chosen), L was set equal to 7 (ie R became the last 7 rounds of the game) and the criterion was computed again. This process ended at $L = \lambda - 1$ when the criterion was, for the first time, not met (given the same value of ε). At that point, the algorithm came to a halt and convergence to a stable pattern of discrimination was pronounced to have occurred on round $32 - \lambda$.

The meaning of the above criterion is that the larger of absolute deviations between

(*a*) the empirical probabilities over the run's L rounds that *A*-players and *D*-players will play strategy h, and
(*b*) the same empirical probabilities as observed *in the last 5 rounds* the smaller the chances that the pattern of discrimination which we observe in the last 5 rounds had 'settled down' L rounds before the game's end. Thus the criterion checks that the larger absolute deviation among (*a*) and (*b*) must *not* exceed a certain threshold ε.

Finally, the value of ε was selected in such a manner that, if the convergence criterion were to hold, then we could be certain with 95% confidence that, in the last L rounds of the game, π' and θ'' had converged to their values in the last 5 rounds. The table below, based on the above algorithm, reports on whether convergence was achieved and if so during which round:

Convergence table

Session no. and colour treatment		Game *HD*			Game *HDC*		
		Convergence?	Which Colour?	Which Round?	Convergence?	Which colour?	Which Round?
9	HDC-HD	Yes	Red	26	No	–	–
10	HD-HDC	Yes	Red	24	Yes	Red	12
11	HD-HDC	Yes	Blue	19	Yes	Blue	8
12	HD-HDC	Yes	Blue	18	Yes	Blue	5
13	HDC-HD	Yes	Blue	1	Yes	Blue	26
14	HDC-HD	Yes	Red	24	No	–	–
15	HD-HDC	Yes	Red	21	Yes	Red	11
16	HD-HDC	Yes	Blue	20	Yes	Blue	2
17	HDC-HD	Yes	Red	23	No	–	–
18	HDC-HD	No	–	–	No	–	–
19	HD–HDC	Yes	Blue	15	Yes	Blue	8
20	HD–HDC	Yes	Blue	20	Yes	Blue	6
21	HDC–HD	No	–	–	No	–	–
22	HD–HDC	Yes	Red	14	Yes	Red	13
23	HD–HDC	Yes	Blue	16	Yes	Blue	1
24	HD–HDC	Yes	Red	–	Yes	Red	2
25	HD–HDC	Yes	Blue	10	Yes	Blue	20
26	HDC–HD	No	–	–	No	–	–
27	HD–HDC	Yes	Red	20	Yes	Red	21
28	HD–HDC	Yes	Red	7	Yes	Red	6
29	HDC–HD	No	–	–	No	–	–
30	HD–HDC	Yes	Blue	18	Yes	Blue	7
31	HD–HDC	No	–	–	No	–	–
32	HD–HDC	Yes	Blue	16	Yes	Blue	10

Appendix C: Disaggregated data from all 32 rounds of treatment HD-HDC-Clr

In this Appendix, we present the data for *all* 32 rounds of *each* game in *HD-HDC-Clr* corresponding to Table 7 (in which only data from the last 11 rounds of HDC were

reported). Italicised figures signify that the relevant observation was different from those in the same column at the 99% confidence level. Note that only data from 15 out of the 16 sessions of *HD-HDC-Clr* were used (since in session 31 – see Appendix B – no colour emerged as 'advantaged').

Data from all 32 rounds of HD in the 15 HD-HDC-Clr sessions in which A and D colours emerged

	Outcomes			Strategies	
	(h, h)	(h, d)	(d, d)	'h'	'd'
Meetings between two *A* players	30.9	44.2	24.9	53	47
Meetings between two *D* players	26.8	43.2	30	48.4	51.6
Meetings between an *A* and a *D* player	*14*	*59.9*	26.1	44	56

Data from all 32 rounds of HDC in the same 15 HD-HDC-Clr sessions as above

	Outcomes						Strategies		
	(h, h)	(h, d)	(d, d)	(c, c)	(h, c)	(d, c)	'h'	'd'	'c'
Meetings between two *A* players	*42.8*	17.2	3	3	24.2	9.8	*63.5*	16.5	20
Meetings between two *D* players	17.7	12.8	6	*29*	*10.3*	*24.3*	29.2	24.5	*46.3*
Meetings between an *A* and a *D* player	8.2	*62.3*	0	2.6	22.7	4.2	50.7	*37.3*	16.1

The next table presents a further breakdown of the above data as it pertains to meetings between an *A* and a *D* player. Note that the data refers to game HDC (with the corresponding data from game HD in brackets). For example, in HDC, there were no occurrences of (d,d) when an *A*-player met a *D*-player, whereas that outcome occurred 26.1% of the time when an *A*-player met a *D*-player in game HD.

Aggregate behaviour in HDC (*HD data in parenthesis*) *when an A-player met a D-player in HD-HDC-Clr*

		D-player			
		'h'	'd'	'c'	*Sub-total*
A-player	'h'	8.2 (14)	62.3 (48.1)	8.5	*79 (62.1)*
	'd'	0 (11.8)	0 (26.1)	1.2	*1.2 (37.9)*
	'c'	14.2	3	2.6	*19.8*
	Sub-total	*22.4 (25.8)*	*65.3 (74.2)*	*12.3*	*100*

Appendix D: Disaggregated data from all 32 rounds of treatment HDC-HD-Clr

This appendix offers three tables equivalent to those of Appendix C, only this time for treatment *HDC-HD-Clr*. Italicised figures again signify that the relevant observation was different from those in the same column at the 99% confidence level. As in Appendix C, note that only data from the sessions of *HDC-HD-Clr* in which discrimination on the basis of colour emerged were used. That is, the data below refer to only 4 out of the 8 sessions of treatment *HDC-HD-Clr* for game HD and only 1 session for game HDC (see Appendix B).

Data from all 32 rounds of HD in the 4 HDC-HD-Clr sessions in which A and D colours emerged

	Outcomes			Strategies	
	(h, h)	(h, d)	(d, d)	'h'	'd'
Meetings between two *A* players	31.7	39.9	28.4	51.6	48.4
Meetings between two *D* players	31.9	36.1	32	49.9	50.1
Meetings between an *A* and a *D* player	21.8	*52.4*	25.8	48	52

Data from all 32 rounds of HDC in the single HDC-HD-Clr session where discrimination surfaced

	Outcomes						Strategies		
	(h, h)	(h, d)	(d, d)	(c, c)	(h, c)	(d, c)	'h'	'd'	'c'
Meetings between two *A* players	32.5	8.9	0.4	7	35.3	15.9	54.6	12.8	32.6
Meetings between two *D* players	32.4	7.2	4.3	10.1	30	16	51	15.9	33.1
Meetings between an *A* and a *D* player	27.8	6.2	1.8	5.8	36.8	21.6	49.3	15.7	35

Aggregate behaviour in HDC (HD data in parenthesis) when an A-player met a D-player in HDC-HD-Clr

		D-player			
		'h'	'd'	'c'	*Sub-total*
A-player	'h'	27.8 (21.8)	3 (31.9)	17	*47.8* (*53.7*)
	'd'	3.2 (20.5)	1.8 (25.8)	11.5	*16.5* (*46.3*)
	'c'	19.8	10.1	5.8	*35.7*
	Sub-total	*50.8* (*42.3*)	*14.9* (*57.7*)	*34.3*	*100*

Appendix E: Payoffs

Overall average pay-offs per player per round

	HD-HDC-NClr	*HDC-HD-NClr*	*HD-HDC-Clr*	*HDC-HD-Clr*
HD	13c	1c	35.8c	19.3c
HDC	48.7c	29.5c	93.3c	74.7c

Pay-offs per player per round in colour sessions where discrimination evolved

Treatment		*HD-HDC-Clr*		*HDC-HD-NClr*	
Pairing	Game	*A*'s pay-off	*D*'s pay-off	*A*'s pay-off	*D*'s pay-off
Meetings between an *A* and a *D* player	HD	88.3c	15.7c	48.6c	69.2c
	HDC	$1.36	39.7c	$1.93	$1.36
Meetings between two *A* players	HD	7.3c	–	$1.18	–
	HDC	16.2c	–	$2.43	–
Meetings between two *D* players	HD	–	19.6c	–	$1.3
	HDC	–	$1.01	–	$2.45
Mean per game	HD	47.8c	17.7c	83.1c	$1
	HDC	75.9c	70.5c	$2.18	$1.91
Overall average		*61.9c*	*44.1c*	*$1.51*	*$1.45*

Appendix F: Rabin's (1993) fairness equilibria in HD and HDC

Rabin (1993) specifies player A's utility function as

$$U_A(s) = v\mathrm{ER}(s) + \psi(s) \tag{1}$$

where ER(s) are the expected monetary returns from strategy s, ψ are *the psychological utility gains* from one's choice, and v is her *marginal monetary valuation* (i.e. the marginal utility of expected material returns, normalised so that the marginal psychological utility derived from playing s equals one). Function $\psi(s)$ is then defined further by

$$\psi(s) = \phi_B(s)[1 + f_A(s)] \tag{2}$$

where f_A is a function which takes a positive (negative) value if A is being kind (unkind) to B (i.e. her opponent) and ϕ_B is another function which takes a positive (negative) value if A *anticipates* kindness (unkindness) from B. Evidently, if A anticipates kindness (unkindness) from B, she loses some utility when acting unkindly (kindly) towards B *ceteris paribus*. In more detail, Rabin postulates A's two (un)kindness functions as

$$f_A(s_A, s_B) = \frac{P^B - E^B}{H^B - L^B} \quad \text{and} \quad \phi_A(s_A, s_B) = \frac{P^A - E^A}{H^A - L^A} \tag{3}$$

where (s_A, s_B) are, respectively, A's strategy and the strategy that A predicts B will be playing; P^A and P^B are the two players' pay-offs from outcome (s_A, s_B); H^B (H^A) is the highest pay-off B (A) can hope to receive if she plays strategy s_B (s_A); L^B (L^A) is the lowest payoff B (A) can hope to receive if she plays strategy s_B (s_A); and E^B (E^A) is the pay-off to which B (A) is *entitled* if she plays strategy s_B (s_A) as *perceived* by A at the moment of choice.

Clearly, if by playing strategy s_A, when anticipating B to choose s_B, A thinks that B will receive less than what B is entitled to (ie if $P^B - E^B < 0$) then A feels she is being unkind to B and her kindness function $f_A(s_A, s_B)$ is negative. Similarly, if A thinks that in playing s_B (while A chooses s_A) B is giving A a pay-off greater than A's is entitlement to (ie if $P^A - E^A > 0$) then A believes that B is being kind; ie $\phi_B(s_A, s_B) > 0$. So Rabin's notion of fairness, as captured in this model, requires that A should be prepared to sacrifice at least some of her expected pay-off to be kind (unkind) to a B whom she expects to be kind (unkind) in return. (Note that if A expects B to be kind, she expects ϕ_B to be positive. In this case, A loses utility (ie $\psi < 0$) if A is unkind to B; that is, if she chooses a strategy such that f_A is negative. The opposite is true if A expects B to be unkind.)

To complete the utility transformation, Rabin needs a definition of what a player is 'entitled' to when she plays a certain strategy. He argues that A's (B's) entitlement from playing strategy s_A (s_B) is the average of her pay-offs corresponding to her best and worst Pareto optimal outcomes given that she chose s_A (s_B). For example, a player playing 'h' in HD is entitled to \$2 (since outcome (2,0) Pareto dominates (−2,−2)) whereas a player in HDC playing 'c' is entitled to \$1 (the average of −1 and 3; ie her pay-offs from the two Pareto undominated outcomes corresponding to 'c' play). It is important to note that, once these entitlements are inserted in the kindness functions above, different second-order beliefs give rise to different utility transformations and thus *different pay-off matrices*. As a result, Rabin can only describe the utility pay-offs *in equilibrium*. In our HD and HDC the games' equilibrium normal form (ie the utility

pay-offs corresponding to an equilibrium between both players' first- and second-order beliefs) is given as follows:

$$\text{HD}: \begin{pmatrix} -2 & 2 \\ 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} -2v & 2v \\ -1/2 & v+3/4 \end{pmatrix}$$

$$\text{HDC}: \begin{pmatrix} -2 & 2 & 4 \\ 0 & 1 & 0 \\ -1 & 0 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} -2v-5/36 & 2v-1/2 & 4v+1/8 \\ -5/12 & v+3/4 & -3/8 \\ -v-7/24 & -1/8 & 3v+3/4 \end{pmatrix}$$

Thus pure strategy Nash equilibrium (h,d) ceases to be an equilibrium as long as v is less than 0.75 in HD and 0.833 in HDC. The condition for (d,d) and (h,h) to become (fairness) equilibria in HD is that v falls below 0.75 and 0.25 respectively. In HDC, (d,d), (c,c) and (h,h) become (fairness) equilibria at values of v below 0.833, 0.625 and 0.153 respectively.

References

Asdigan, N., Cohn, E. and Blum, M. (1994). 'Gender differences in distributive justice: the role of self-representation revisited', *Sex Roles*, vol. 30, pp. 303–18.

Babcock, L., Lowenstein, G., Issachoroff, S. and Camerer, C. (1995). 'Biased judgments of fairness in bargaining', *American Economic Review*, vol. 85, pp. 1337–43.

Bacharach, M. (1997). '*We* equilibria: a variable frame theory of co-operation', mimeo.

Bacharach, M. (1999). 'Interactive team reasoning: a contribution to the theory of co-operation', *Research in Economics*, vol. 53, pp. 117–47.

Bacharach, M. and Becnasconi, M. (1997) 'The variable frame theory of focal points: an experimental study', *Games and Economic Behavior*, vol. 19, pp. 1–45.

Binmore, K. and Samuelson, L. (1993). 'Learning to play the ultimatum game', mimeo.

Camerer, C. and Thaler, H. (1995). 'Anomalies: ultimatum, dictators and manners', *Journal of Economic Perspectives*, vol. 9, pp. 209–19.

Erev, I. and Roth, A. (1998). 'Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria', *American Economic Review*, vol. 88, pp. 848–81.

Erev, I., Bereby-Meyer, Y. and Roth, A. (1999). 'The effect of adding a constant to all payoffs: experimental investigation, and implications for reinforcement learning models', *Journal of Economic Behavior & Organization*, vol. 39, pp. 111–28.

Friedman, D. (1996). 'Equilibrium in evolutionary games: some experimental results', ECONOMIC JOURNAL, vol. 106, pp. 1–25.

Hume, D. (1740, 1888). *Treatise of Human Nature*, (L. A. Selby-Bigge, ed.), Oxford: Oxford University Press.

Kahn, A., O'Leary, V., Krulewitz, J. and Lamm, H. (1980). 'Equity and equality: male and female means to a just end', *Basic and Applied Social Psychology*, vol. 1, pp. 173–97.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

Major, B. and Adams, J. (1983). 'Role of gender, inter-personal orientation and self-representation in distributive justice behavior', *Journal of Personality and Social Psychology*, vol. 45, pp. 598–608.

Major, B., Bylsma, W. and Cozzarelli, C. (1989). 'Gender differences in distributive justice preferences: the impact of domain', *Sex Roles*, vol. 21, pp. 487–97.

McDaniel, T., Rutström, E. and Williams, M. (1994). 'Incorporating fairness into game theory and economics: an experimental test with incentive compatible belief elicitation', mimeo.

Mehta, J., Starmer, C. and Sugden, R. (1994). 'The nature of salience: an experimental investigation of pure coordination games', *American Economic Review*, vol. 84, pp. 658–73.

Rabin, M. (1993). 'Incorporating fairness into economics and game theory', *American Economic Review*, vol. 83, pp. 1281–302.

Schotter, A., Weiss, A. and Zapater, I. (1996). 'Fairness and survival in ultimatum and dictatorship games', *Journal of Economic Behavior & Organization*, vol. 31, pp. 37–56.

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

Sugden, R. (2000). 'The motivating power of expectations', in (J. Nida-Rümelin and W. Spohn, eds.), *Rationality, Rules and Structure*, Amsterdam: Kluwer, pp. 103-29.

Varoufakis, Y. (1996). 'Moralising in the face of strategic weakness: experimental clues for an ancient puzzle', *Erkenntnis*, vol. 46, pp. 87–110.

Weibull, J. (1995). *Evolutionary Game Theory*. Cambridge, MA: MIT Press.