ELSEVIER

BioSystems 37 (1996) 135–145

Evolving continuous behaviors in the Iterated Prisoner's Dilemma

Paul G. Harrald*[a], David B. Fogel[b]

[a]*Manchester School of Management, UMIST, P.O. Box 88, Manchester, UK*
[b]*Natural Selection, Inc., 1591 Calle De Cinco, La Jolla, CA 92037, USA*

Abstract

Evolutionary programming experiments are conducted on a variant of the Iterated Prisoner's Dilemma. Rather than assume each player having two alternative moves in the stage-game, cooperate or defect, a continuum of possible moves are available. Players' strategies are represented by feed-forward perceptrons with a single hidden layer. The population size and the number of nodes in the hidden layer are varied across a series of experiments. The results of the simulations indicate a minimum amount of complexity is required in a player's strategy in order for cooperation to evolve. Moreover, under the evolutionary dynamics of the simulation, cooperation does not appear to be a stable outcome.

Keywords: Evolutionary programming; Continuous behaviours; Iterated Prisoner's Dilemma

1. Introduction

The Iterated Prisoner's Dilemma (IPD) has become a standard model for abstracting relationships between individuals of a social group and for studying condition that would foster the evolution of cooperative or selfish behavior within such a community. The game formally involves the time-sequential interaction of two individuals, each of which can adopt one of two moves in each iteration: cooperate or defect. *Cooperation* (C) implies increasing the reward of both participants, while *defecting* (D) implies increasing one's own reward at the expense of the other player. The general form of the one-shot prisoner's dilemma (PD) game played at each iteration is presented in Table 1. The payoff matrix that defines the one-shot game adheres to the following constraints:

$$2\gamma_1 > \gamma_2 + \gamma_3,$$
$$\gamma_3 > \gamma_1 > \gamma_4 > \gamma_2.$$

The first constraint removes any obvious desire for mutual cooperation, and also ensures that the payoffs to a series of mutual cooperations is greater than a sequence of alternating plays of

* Corresponding author.

SSDI 0303-2647(95)01550-5

Table 1
The general form of the payoff function in the prisoner's dilemma γ_1 is the payoff to each player for mutual cooperation. γ_2 is the payoff for cooperating when the other player defects. γ_3 is the payoff for defecting when the other player cooperates. γ_4 is the payoff for mutual defection. An entry (α,β) indicates payoffs to players A and B respectively.

		Player B	
		Cooperate	Defect
Player A	Cooperate	(γ_1, γ_1)	(γ_2, γ_3)
	Defect	(γ_3, γ_2)	(γ_4, γ_4)

cooperate-defect against defect-cooperate (which would represent a more sophisticated form of cooperation, see Angeline, 1994). The second constraint ensures that defection is a dominant action, and also that the payoffs accruing to mutual cooperators are greater than those accruing to mutual defectors. In game-theoretic terms, this one-shot game has a single dominant strategy Nash equilibrium, (D,D), which is Pareto dominated by (C,C). This basic irony of the game has been used to characterize *many* instances of economic and biological interaction, from competing oligopolists (Marks, 1989) to the simultaneous hermaphrodite sea-slug *Navanax inermis* (Leonard, 1990).

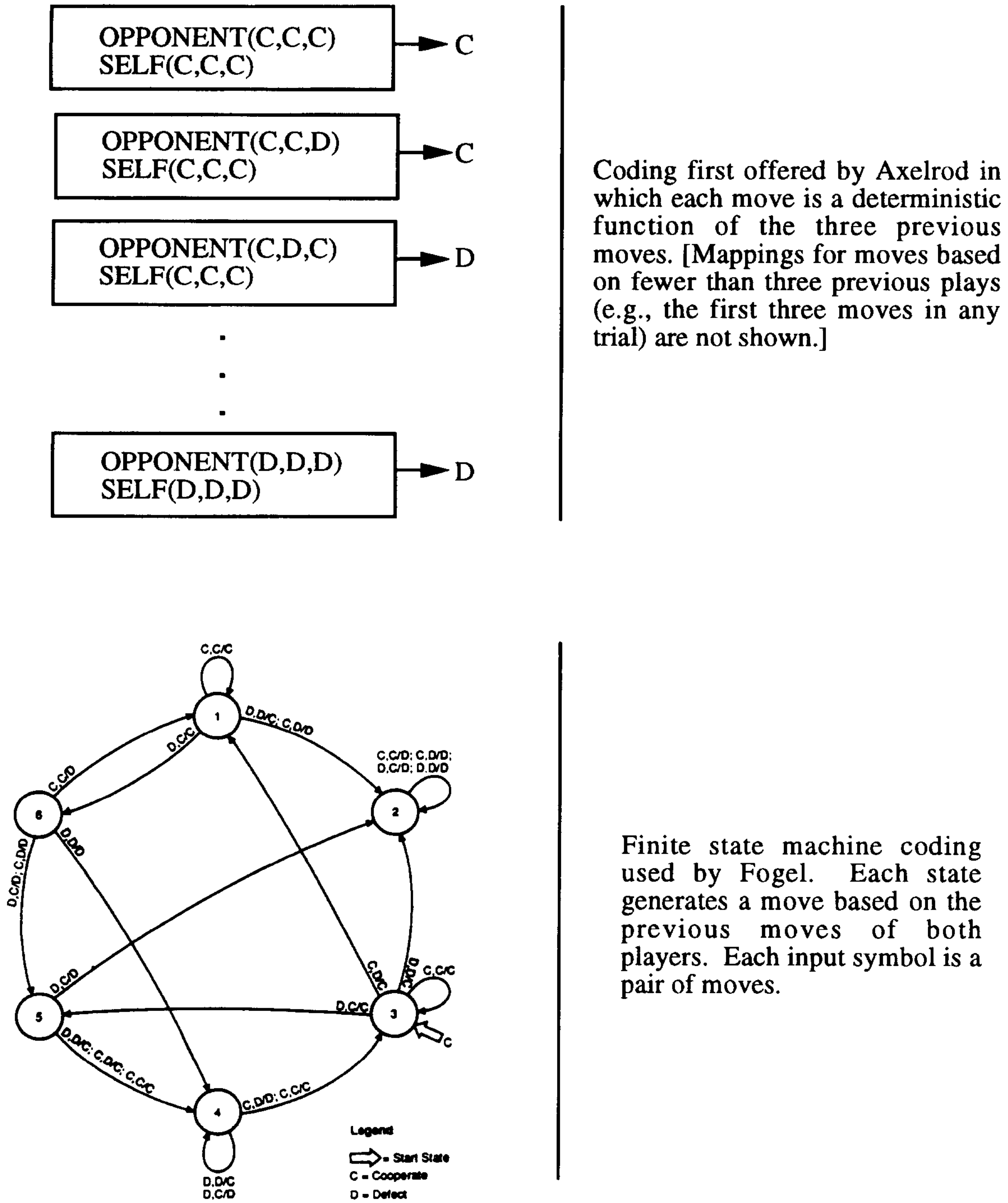
If the game is played for only a single iteration, defection is the only rational move. Defection is also rational if the game is iterated over a series of plays under conditions in which both players' decisions are not affected by previous plays; the game degenerates into a series of independent single trials. But if the players' strategies can depend on the results of previous iterations (i.e., the players adopt 'closed-loop strategies'), 'always defect' is not a dominant strategy: consider a player who will cooperate for as long as his opponent, but should his opponent defect, will himself defect forever. If the game is played for enough iterations, it would be foolish to defect against such a player, at least early on. Thus cooperation can potentially emerge (Kreps et al., 1986). If the game is played infinitely with discounting, game-theoretic analysis stumbles at the Folk Theorem: any sequence of actions can be rationalised as part of an (sub-game perfect) equilibrium (Fundenberg and Maskin, 1986).

In reaction to this, and for many other reasons, much recent research into the IPD has taken its cue from the work of political scientist Robert Axelrod, which analyses the evolution of behavior in the IPD by imposing an evolutionary dynamic on a population of players.

Several such studies have been made (a survey can be found in Harrald, 1995). Such simulations require each player's strategy to be represented by a coding structure. These have taken the form of deterministic mappings on the outcomes of previous interactions (Axelrod, 1987), as well as finite state machines (Fogel, 1993), see Fig. 1. The structures are evolved either by a specific replicator dynamic (Lindgren, 1991), or by implementing evolutionary algorithms such as genetic algorithms (Miller, 1989; Marks, 1989) or evolutionary programming (Fogel, 1991, 1993). Under a traditional payoff matrix used in original tournaments by Axelrod (1980), see Table 2, the typical result of these simulations is for a population to quickly converge toward essentially stable cooperation (Fig. 2). The result appears to be independent of the chosen representation of each player's strategy (Axelrod, 1987; Fogel, 1993). However, there is other work indicating somewhat less stability (Lindgren, 1991; Harrald, 1995).

Two open questions pertaining to the dynamics of the prisoner's dilemma are addressed by the current research: the ability for cooperative behavior to emerge when degrees of cooperation and defection are allowed, and a more general inquiry into whether or not a minimum complexity in players' strategies is required to support cooperation.

In typical simulations of the IPD, players have but two choices, which lie at extremes of the possible spectrum of cooperative behavior. Simply put, if an agent is not defecting, then that agent must be 'fully' cooperating. This severely restricts the range of possible behaviors representable, and does not allow intermediate activity designed to engender cooperation without significant risk, or intermediate behavior designed to quietly or



Coding first offered by Axelrod in which each move is a deterministic function of the three previous moves. [Mappings for moves based on fewer than three previous plays (e.g., the first three moves in any trial) are not shown.]

Finite state machine coding used by Fogel. Each state generates a move based on the previous moves of both players. Each input symbol is a pair of moves.

Fig. 1. Two different forms for representing players' strategies. Deterministic mappings on combinations of previous moves have been employed (Lindgren, 1991; Axelrod, 1987). A specific combination of as many as three previous pairs of plays determined the next move. Finite state machines of up to eight states have been employed (Miller, 1989; Fogel, 1993), in which the current state and the previous pair of plays determine the next state.

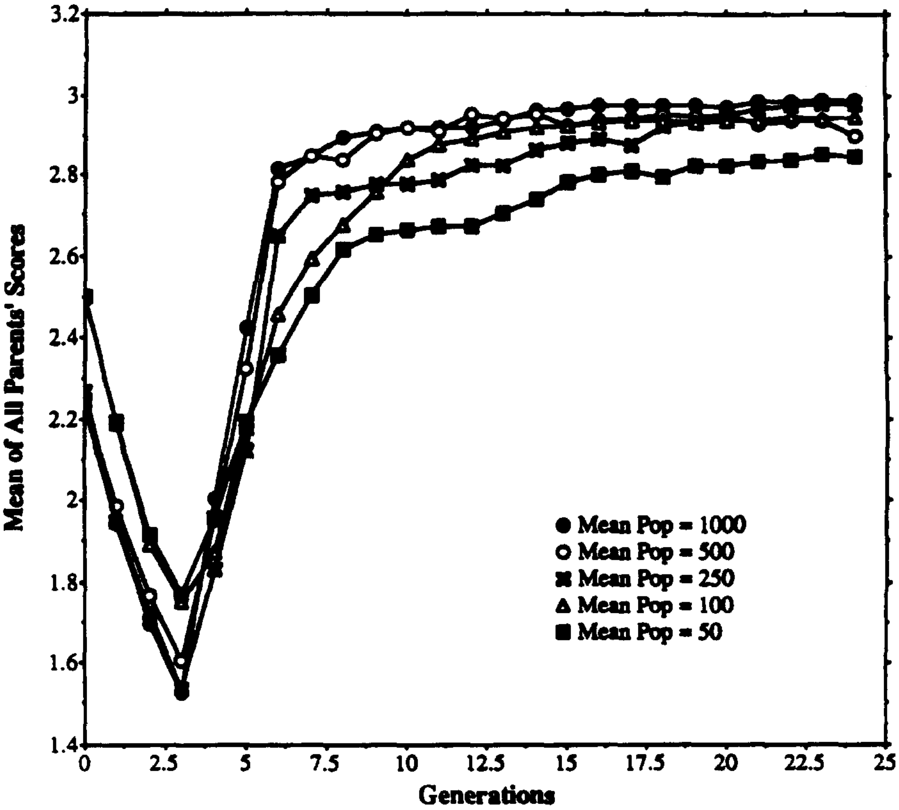


Fig. 2. A typical result of the mean payoff to all parents in an evolving population of finite state machines when faced with the payoff matrix shown in Table 2 (Fogel, 1993) with population size varying from 50 to 1000 parents. The population's mean drops initially, indicating a propensity for players to defect and the gradual eradication of pathological cooperators, but then rapidly rebounds toward mutual coöperation. Similar behavior is described in Axelrod (1987).

surreptitiously take advantage of a partner (To, 1988; Marks, 1989). Hence, once behaviors evolve that cannot be fully taken advantage of (those that punish defection), such strategies enjoy the full and mutual benefits of harmonious cooperation. Naturally, strategies that are able to punish, and also are able to effectively discriminate between likely partners for mutual cooperation and more opportunistic types, must have a sufficient structure to be able to process information and form such conclusions (or behave as-if they did). Once behavioral choices become more complex, as in the present case, it seems unlikely that a simple structure can be so discriminating. The task at hand is to provide a game in the spirit of the PD to evolve a population coded so that there is a meaningful interpretation of 'complexity' in a game for which there are degrees of cooperative behavior.

Table 2
A specific payoff function (Axelrod, 1980)

		Player *B*	
		Cooperate	Defect
Player *A*	Cooperate	(3,3)	(0,5)
	Defect	(5,0)	(1,1)

2. Experimental procedure and results

A strategy in the IPD can be viewed as a transfer function: a system that operates on a sequence of observed moves from previous plays of the stage game and then generates a move for the current play. Multi-layer feed-forward perceptrons (MLPs) provide a convenient form for achieving the required mapping function. Specifically, each player's strategy was represented by a MLP that possessed six input nodes, a prescribed number of hidden nodes, and a single output node. The first three inputs corresponded to the previous three moves of the opponent, while the second three corresponded to the previous three moves of the network itself (Fig. 3). The behavior on any move was described by the continuous range $(-1,1)$, where -1 represents complete defection and 1 represents complete cooperation. All nodes in the network used sigmoidal filters scaled to yield output between -1 and 1. The output of the network was taken as its move in the current iteration.

A traditional payoff matrix (Axelrod, 1987) was approximated by a planar equation of both players' moves. Specifically, the payoff to player A against a player B was given by:

$$f(\alpha,\beta) = -0.75\alpha + 1.75\beta + 2.25$$

where α and β are the moves of the players A and B respectively. This function is shown in Fig. 4. The basic tenor of the one-shot PD is maintained: full defection is the dominant move, joint payoffs are maximised by mutual full cooperation.

The evolutionary program was implemented as follows:

1. A population of a given number of networks was initialized at random. All of the weights and biases of each network were initialized uniformly over $[-0.5,0.5]$.
2. A single offspring network was created from each parent by adding a standard Gaussian random variable to every weight and bias term.
3. All networks played against each other in a round-robin competition (each met every other one time). Encounters lasted 151 moves and the fitness of each network was assigned according to the average payoff per move.

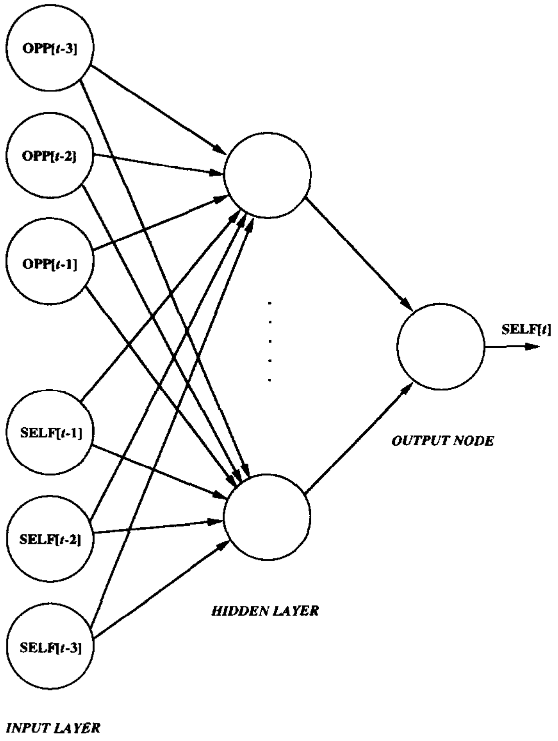


Fig. 3. The neural network architecture used to represent players' strategies in the current experiments. The six input nodes correspond to the three previous moves of each player. In the figure, OPP[x] represents the opponent's move at time x, while SELF[x] represents the network's own move at time x. The output of the network is the play at time t (the current time). The hidden and output nodes perform weighted sums of the inputs offset by a variable bias and pass this result, denoted by b, through a nonlinear sigmoid filter of the form $2[(1 + \exp(-b))^{-1} - 0.5]$, yielding a value ranging over $(-1,1)$. Complete defection is represented by -1, while complete cooperation is represented by 1. When an input node refers to a move that has not yet been played (e.g., on the first iteration of the game, $t = 1$, when there are no previous moves by either player), a value of 0.0 is presented.

4. All networks were ranked according to fitness, and the top half were selected to become parents of the next generation.
5. If the preset maximum number of generations, in this case 500, was met, the procedure was halted; otherwise proceed to step 2.

Two sets of experiments were conducted with various population sizes. In the first, each net-

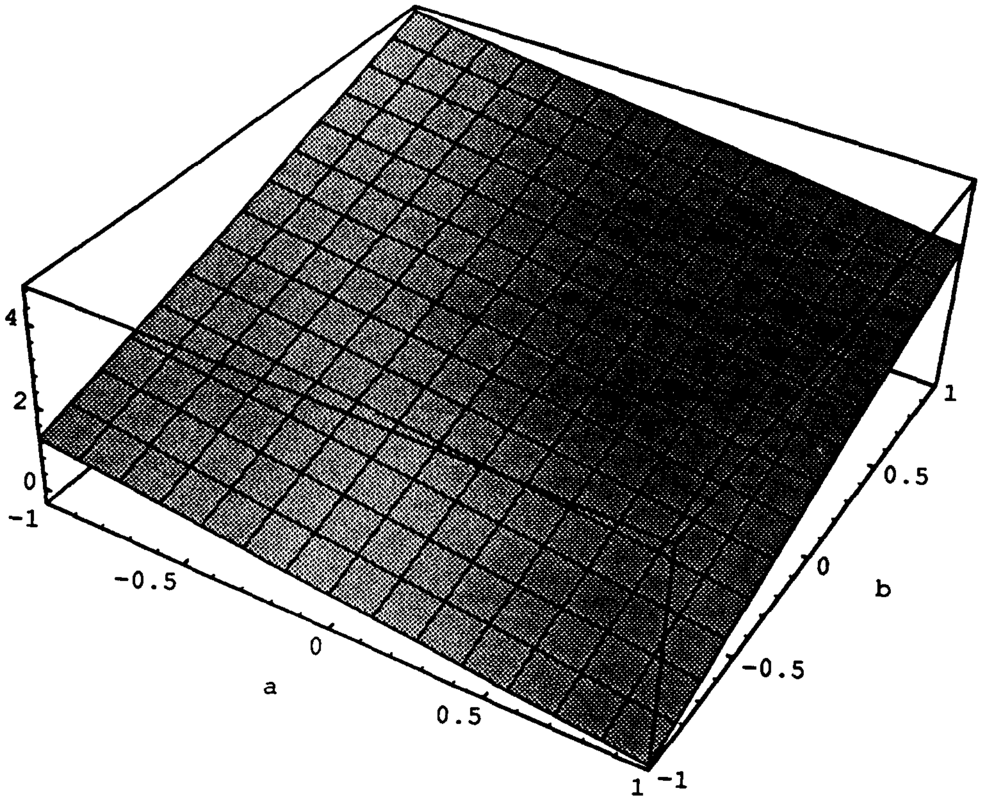


Fig. 4. The planar approximation to the payoff function used in Axelrod (1987). The maximum possible payoff of 4.75 of obtained when player A defects completely, and player B cooperates completely.

work possessed only two hidden nodes (denoted as 6-2-1, for the six input nodes, two hidden nodes, and one output node). This architecture was selected because it is the minimum amount of complexity that requires a hidden layer. In the second, the number of hidden nodes was increased by an order of magnitude to 20. Twenty trials were conducted in each setting with population sizes of 10, 20, 30, 40, and 50 parents.

Table 3 provides results in five behavioral categories. The assessment of apparent trends or instability was admittedly subjective, and in some cases the correct decision was not obvious (e.g., Fig. 5a). But, in general, the results showed: (1) there was no tendency for cooperative behavior when using a 6-2-1 network regardless of population size, (2) above some minimum population size, cooperation was likely when using a 6-20-1 network, (3) any cooperative behavior that did

Table 3

Tabulated results of the 20 trials in each setting. The columns represent: (a) the number of trials that generated cooperative behavior after the 10th generation, (b) the number of trials that demonstrated a trend toward increasing mean payoffs, (c) the number of trials that demonstrated a trend toward decreasing mean payoffs, (d) the number of trials that generated persistent universal complete defection after the 200th generation, and (e) the number of trials that appeared to consistently generate some level of cooperative behavior.

6-2-1	(a)	(b)	(c)	(d)	(e)
10 Parents	0	0	10	9	0
20 Parents	6	0	19	13	0
30 Parents	4	1	19	11	0
40 Parents	7	0	19	12	0
50 Parents	2	0	10	2	0
6-20-1	(a)	(b)	(c)	(d)	(e)
10 Parents	9	2	13	11	4
20 Parents	16	5	10	3	15
30 Parents	13	2	15	6	13
40 Parents	15	5	14	5	15
50 Parents	15	1	16	2	15

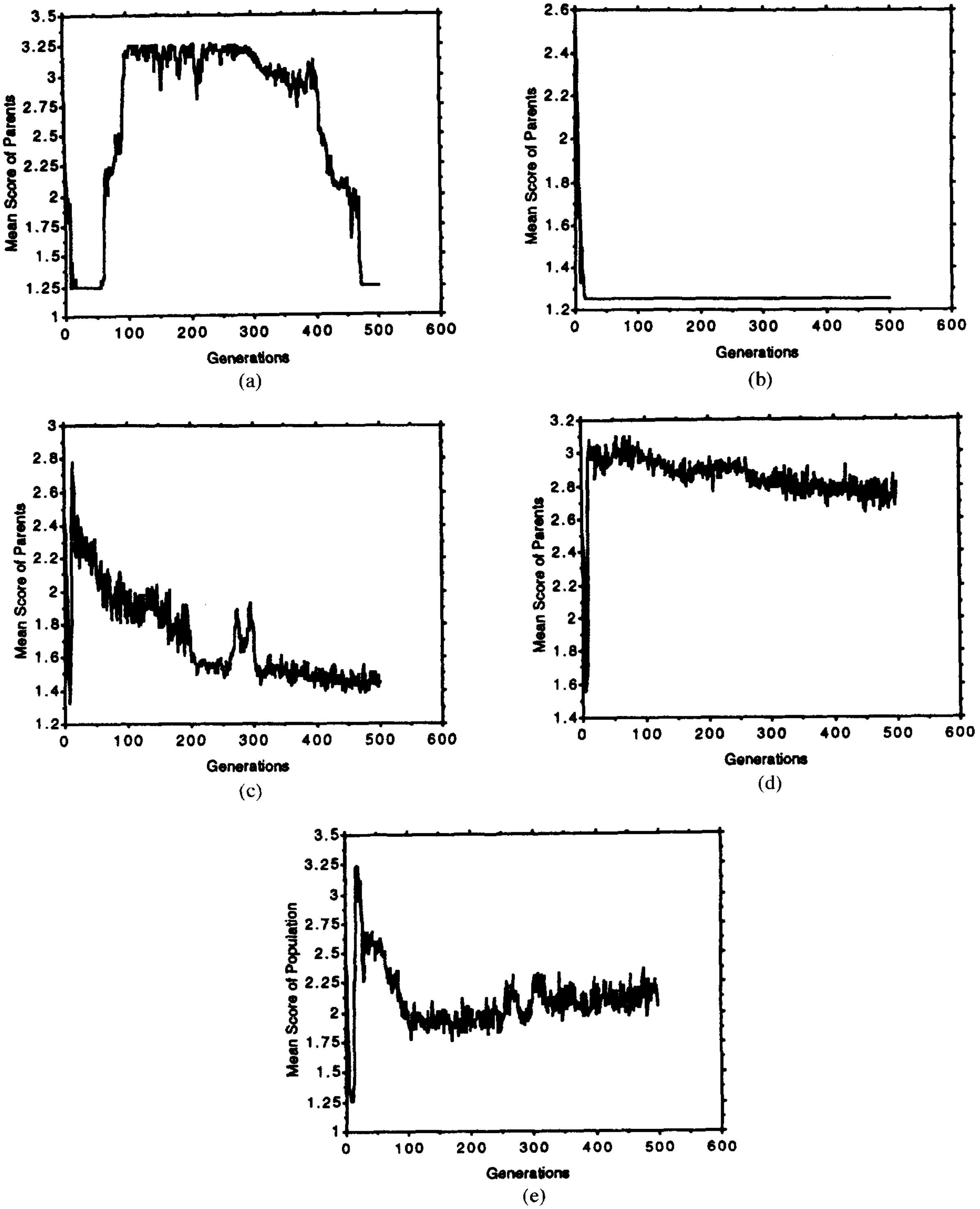


Fig. 5. Examples of various behaviors generated based on the conditions of the experiments. (a) Oscillatory behavior (10 parents, trial #4 with 6-2-1 networks), (b) complete defection (30 parents, trial #10 with 6-20-1 networks), (c) decreasing payoffs leading to further defection (30 parents, trial #9 with 6-2-1 networks), (d) general cooperation with decreasing payoffs (50 parents, trial #4 with 6-20-1 networks), (e) an increasing trend in mean payoff (30 parents, trial #2 with 6-2-1 networks).

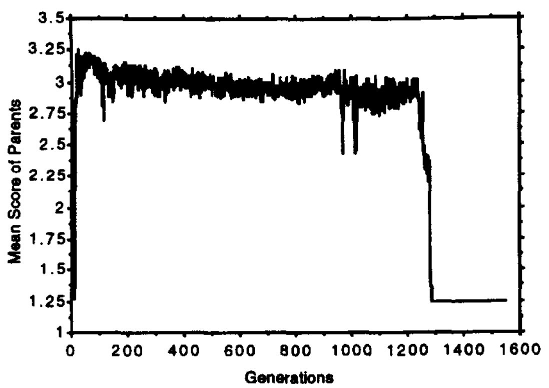


Fig. 6. The result of trial #10, with 20 parents using 6-20-1 networks iterated over 1500 generations. The population's behavior changes rapidly toward complete defection after the 1200th generation.

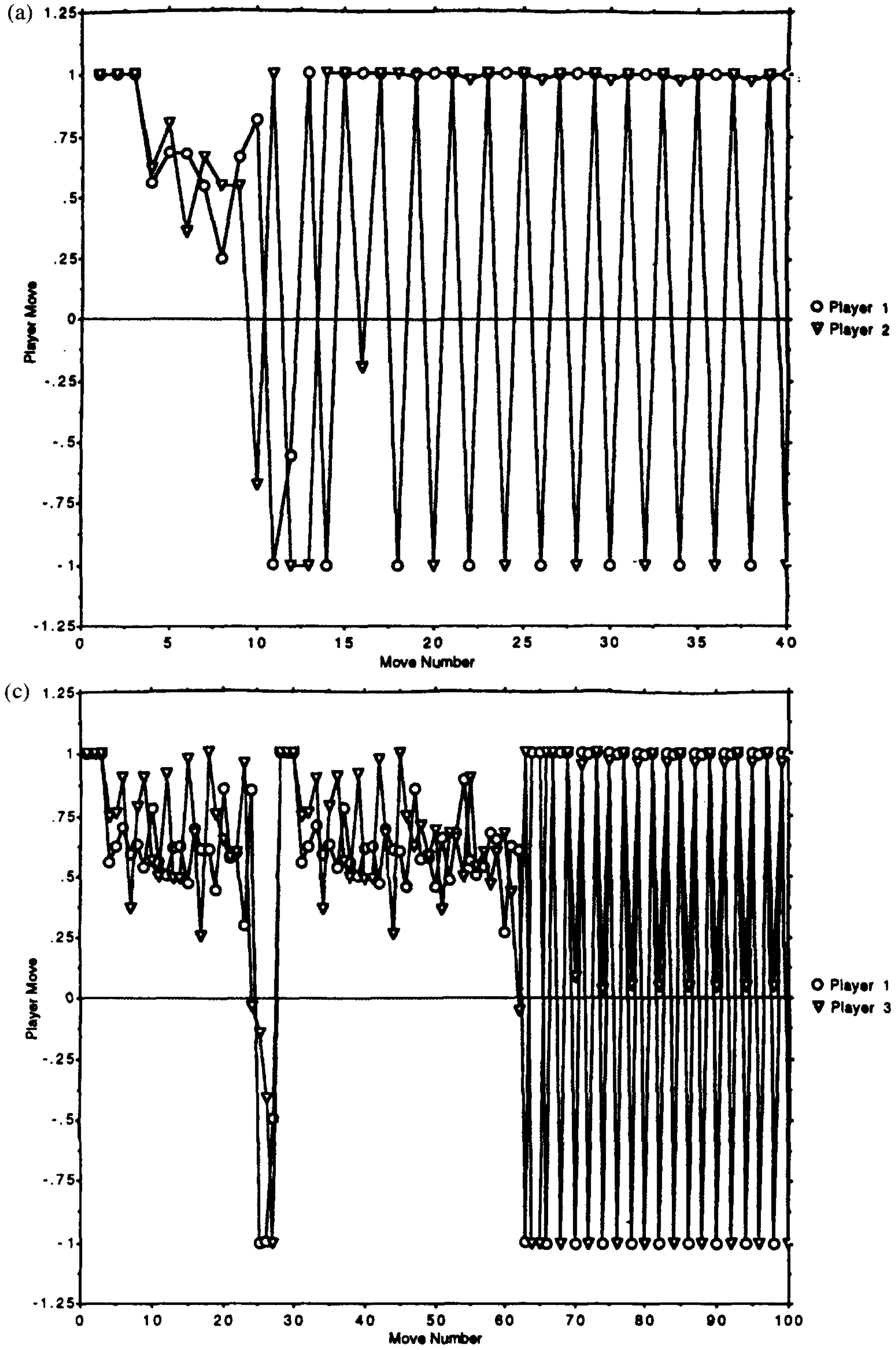
arise did not tend toward complete cooperation, and (4) complete and generally unrecoverable defection was the likely result with the 6-2-1 network, but could occur even when using 6-20-1 networks. Fig. 5 indicates some of the typical observed behavioral patterns.

3. Conclusions

The results suggest that the evolution of mutual cooperation in light of the chosen payoff function and continuous behaviors requires a minimum complexity (in terms of behavioral freedom) in the policies of the players. A single hidden layer perceptron is capable of performing universal function approximation if given sufficient nodes in the hidden layer. Thus, the structures used to represent player policies in the current study could be tailored to be essentially equivalent to the codings in which each move was a deterministic function of the previous three plays of the game (Axelrod, 1987). While previous studies observed stable mutual cooperation (Axelrod, 1987), cooperative behavior was never observed with the 6-2-1 networks, but was fairly persistent with the 6-20-1 networks. But the *level* of cooperation that was generated when using the 6-20-1 networks was neither complete nor steady. Rather, the mean payoff to all parents tended to peak below a value of 3.0 and decline, while complete mutual cooperation would have yielded an average payoff of 3.25.

The tendency for cooperation to evolve with sufficient complexity should be viewed with caution, for at least three reasons. First, very few trials with 6-20-1 networks exhibited an increase in payoff as a function of the number of generations. The more usual result was a steady decline in mean payoff, away from increased cooperation. Second, cooperative behavior was not always steady. Fig. 6 indicates the results for trial 10 with 20 parents using 6-20-1 networks executed over 1500 generations. The behavior appeared cooperative until just after the 1200th generation, at which point it declined rapidly to a state of complete defection. A recovery from complete defection was rare, regardless of the population size or complexity of the networks. It remains to be seen if further behavioral complexity (i.e., a greater number of hidden nodes) would result in a more stable cooperation. Finally, the specific level of complexity in the network that must be attained before cooperative behavior emerges is not known, and there is no *a priori* reason to believe that there is a smooth relationship between the propensity to generate cooperation and strategic complexity.

It was of interest to identify whether or not the evolving networks were generating moves that were close to complete cooperation or defection, or instead tending toward more intermediate forms. Fig. 7 indicates the sequences of moves from the best evolved network when playing the second, third, and fourth highest scoring networks after 500 generations in the 10th trial with 20 parents using 6-20-1 architectures. There was a repeated pattern of initial complete mutual cooperation, but this quickly degenerated into cyclic behavior with moves covering the range from complete cooperation to complete defection. In the case of the best network playing the fourth best, the latter offered complete cooperation, but the best generated a behavior of 0.663, thereby taking advantage of its opponent. In general, there was no evidence of an elimination of intermediate behaviors.


(a)
1.25
1
.75
.5
.25
0
-.25
-.5
-.75
-1
-1.25
Player Move
0
5
10
15
20
25
30
35
40
Move Number
○ Player 1
▽ Player 2
(c)
1.25
1
.75
.5
.25
0
-.25
-.5
-.75
-1
-1.25
Player Move
0
10
20
30
40
50
60
70
80
90
100
Move Number
○ Player 1
▽ Player 3

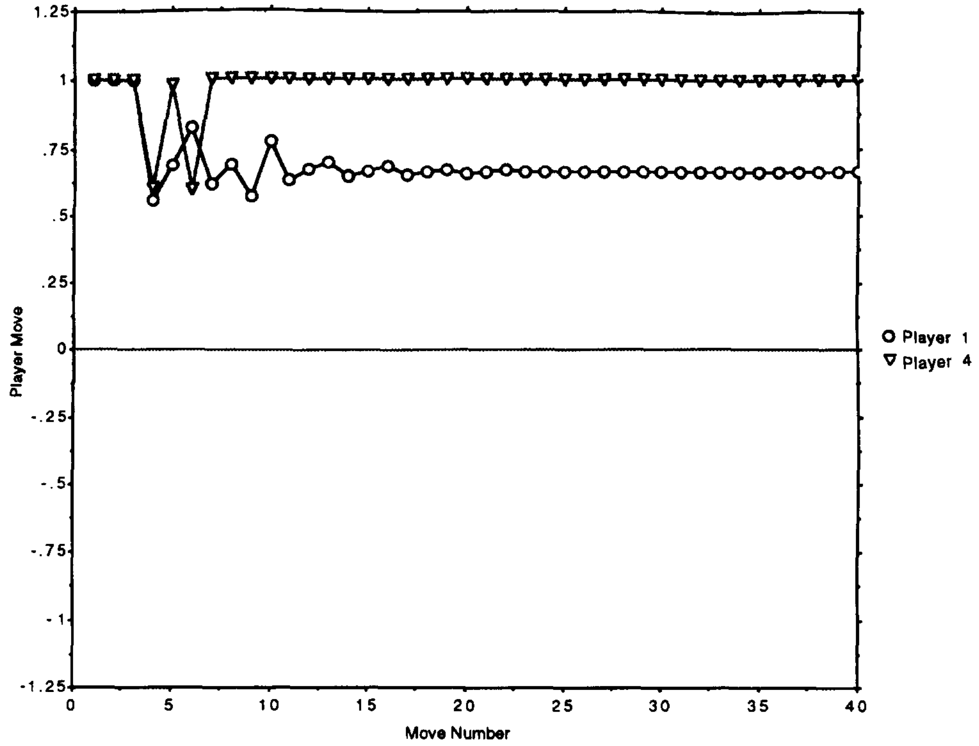


Fig. 7. Sequences of moves generated by the best evolved network at the 500th generation in the 10th trial of 20 parent 6-20-1 networks when playing the (a) second, (b) third, and (c) fourth best networks. The evidence indicates repeated oscillations of cooperation and defection, and no tendency to eliminate intermediate behaviors.

The population size did appear to influence the probability of generating cooperative behavior. When utilizing 6-20-1 networks and a population of 10 parents, only one trial out of 10 consistently generated cooperative behavior. Yet the majority of trials with larger populations did generate cooperatiave behaviour. These results must be interpreted with care, however, because rapid jumps were observed in the mean payoff of the population, even when using 30 or more parents. The mean behavior of the surviving parents sometime fluctuated between nearly complete cooperation to near neutrality or nearly complete defection within one or only a few generations. Whether this was a result of parents interacting with diverse offspring, or simply parents being overthrown by their offspring remains a subject for future study.

Acknowledgements

The authors would like to thank D. Andolfatto, W. Atmar, G.B. Fogel and L.J. Fogel for their comments and criticism.

References

Angeline, P.J., 1994. An alternative interpretation of the iterated prisoner's dilemma and the evolution of non-mutual cooperation, in: Artificial Life IV, R.A. Brooks and P. Maes (eds.) (MIT Press, Cambridge, MA), pp. 353-358.

Axelrod, R., 1980. Effective choice in the prisoner's dilemma. J. Conf. Resol. 24, 3-25.

Axelrod, R., 1987. The evolution of strategies in the iterated prisoner's dilemma, in: Genetic Algorithms and Simulated Annealing, L. Davis (ed.) (Pitman, London), pp. 32-41.

Fogel, D.B., 1991. The evolution of intelligent decision making in gaming. Cybern. Sys. 22, 223-236.

Fogel, D.B., 1993. Evolving behavior in the iterated prisoner's dilemma. Evol. Comput. 1:1, 77-97.

Fundenberg, D. and Maskin, E., 1986. The folk theorem for repeated games with discounting or with incomplete information. Econometrica 56, 533–544.

Harrald, P.G. 1995. Evolving behaviours in repeated games 2-player, in: Practical Handbook of Genetic Algorithms, L. Chambers (ed.) (CRC Publishers, Boca Raton, FL).

Kreps, D., Milgrom, P., Roberts, J. and Wilson, J., 1982. Rational cooperation in the finitely repeated prisoner's dilemma. J. Econ. Theory 27:2, 326–355.

Leonard, J., 1990. The hermaphrodite's dilemma. J. Theor. Biol. 147, 362–372.

Lindgren, K., 1992. Evolutionary phenomena in a simple dynamic, in: Artificial Life II, C.G. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen (eds.) (Addison-Wesley, Reading, MA), pp. 295–312.

Marks, R., 1989. Breeding hybrid strategies: Optimal behavior for oligopolists in: Proceedings of the Third International Conference on Genetic Algorithms. J.D. Schaffer (ed.) (Morgan Kaufmann, San Mateo, CA).

Miller, J.H., 1989. The coevolution of automata in the repeated prisoner's dilemma. Santa Fe Institute Working Paper 89-003.

To, T., 1988. More realism in the prisoner's dilemma. J. Conf. Resol. 32, 402–408.