Multiple Choices and Reputation in Multiagent Interactions

Siang Yew Chong, *Member, IEEE*, and Xin Yao, *Fellow, IEEE*

Abstract—Coevolutionary learning provides a framework for modeling more realistic iterated prisoner's dilemma (IPD) interactions and to study conditions of how and why certain behaviors (e.g., cooperation) in a complex environment can be learned through an adaptation process guided by strategic interactions. The coevolutionary learning of cooperative behaviors can be attributed to the mechanism of direct reciprocity (e.g., repeated encounters). However, for the more complex IPD game with more choices, it is unknown precisely why the mechanism of direct reciprocity is less effective in promoting the learning of cooperative behaviors. Here, our study suggests that the evolution of defection may be a result of strategies effectively having more opportunities to exploit others when there are more choices. We note that strategies are less able to resolve the intention of an intermediate choice, e.g., whether it is a signal to engender further cooperation or a subtle exploitation. A likely consequence is strategies adapting to lower cooperation plays that offer higher payoffs in the short-term view when they cannot resolve the intention of opponents. However, cooperation in complex human interactions may also involve indirect interactions rather than direct interactions only. Following this, we study the coevolutionary learning of IPD with more choices and reputation. Here, current behavioral interactions depend not only on choices made in previous moves (direct interactions), but also choices made in past interactions that are reflected by their reputation scores (indirect interactions). The coevolutionary learning of cooperative behaviors is possible in the IPD with more choices when strategies use reputation as a mechanism to estimate behaviors of future partners and to elicit mutual cooperation play right from the start of interactions. In addition, we study the impact of the accuracy of reputation estimation in reflecting strategy behaviors of different implementations and why it is important for the evolution of cooperation. We show that the accuracy is related to how memory of games from previous generations is incorporated to calculate reputation scores and how frequently reputation scores are updated.

Index Terms—Coevolutionary learning, evolutionary computation, evolutionary games, iterated prisoner's dilemma (IPD), reputation.

I. INTRODUCTION

IN THE iterated prisoner's dilemma (IPD), two players engaged in repeated interactions are given two choices, cooperate and defect [1]–[3]. The dilemma, i.e., both players

who are better off mutually cooperating than mutually defecting are vulnerable to exploitation by one of the party who defects, embodies the tension between rationality of individuals who are tempted to defect and rationality of the group where every individual is rewarded by mutual cooperation [2]. Although the IPD has become a popular model to study conditions for cooperation to occur among selfish individuals, it has also received much attention in many other areas of study, and has been used to model social, economic, and biological interactions [3].

Defection is not necessarily the best choice of play in the IPD. Many studies have shown cooperative play to be a viable strategy [1], [2], which can be learned from an initial, random population using evolutionary algorithms [4]–[8]. The evolution of cooperation can persist for a long period [9] even though it is unstable [10]. Although both the coevolutionary learning approach and the classical evolutionary games approach that considers frequency dependent reproductions of predetermined strategies (e.g., “ecological approach” [2], [11]) can be used as tools for study, the former approach has the advantage of providing a framework for the learning of effective strategies through an adaptation process guided by strategic interactions.

Most of the early studies using the coevolutionary learning approach [4]–[9], [12] considered two choices, i.e., cooperate and defect. However, further studies extended the classical two-choice IPD to more complex interactions that involved more choices (e.g., modeled with continuous levels of cooperation in [13], and multiple, discrete levels of cooperation in [14]–[16]). In these studies [13]–[16], coevolution was used to evolve a population of neural networks in the IPD with more choices.

In general, the studies concluded that cooperation can still be evolved in the IPD with more choices, although the evolution of cooperation is unstable [13] and more difficult to achieve compared with the classical IPD game [14]–[16]. Other studies that used different models, such as the continuous prisoner's dilemma [17] and the spatial prisoner's dilemma [18] (both considered strategies with linear responses to previous moves), also showed that cooperative outcomes are less likely. This raises the question as to why coevolutionary learning models are rather unsuccessful in evolving cooperative play in the IPD with more choices. Regardless of the available number of choices to play in the IPD, strategies still engage in repeated interactions. As in the case of the classical IPD with two choices, cooperative behaviors can emerge among the population of strategies in the IPD with more choices by reciprocating cooperation during their interactions.

Previous studies have shown that certain aspects in coevolutionary learning models can affect the evolutionary outcome. In particular, these studies argued that certain conditions have

Manuscript received January 24, 2006; revised September 18, 2006. The work of X. Yao was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant GR/T10671/01.

S. Y. Chong is with the School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K. (e-mail: S.Y.Chong@cs.bham.ac.uk).

X. Yao is with The Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K. (e-mail: X.Yao@cs.bham.ac.uk).

Digital Object Identifier 10.1109/TEVC.2006.890544

to be met in the coevolutionary learning models to evolve cooperation. For example, Harrald and Fogel [13] argued that a minimum complexity level in the strategy representation is required to evolve cooperative play. Darwen and Yao [14]–[16] argued that behavioral diversity of strategies in the population, not genetic diversity, is the important factor for the evolution of cooperation. We recently showed that introducing behavioral variations in the population is also dependent on the choice of strategy representations because some representations are easier to introduce more behavioral variations [19].

However, it is known that cooperation can be achieved in complex human interactions that are not limited to direct interactions only [20]. This may explain the limitation with the current coevolutionary learning framework for IPD that only involves direct interactions. With a coevolutionary learning model, cooperative behaviors are more difficult to learn for the IPD with more choices [13]–[16], which strongly suggests that the mechanism of direct reciprocity (e.g., repeated encounters) is less effective in promoting cooperation when the number of choices increases.

Given this problem, we investigate one solution that uses the idea that cooperation between two players depends on prior interactions with other players, i.e., indirect reciprocity, where the mechanism for evolving cooperative behaviors is reputation [21]. Here, we incorporate reputation into the existing coevolutionary learning model for IPD games. This is different from previous studies where strategies consider only previous moves as inputs [14]–[16]. For our study, strategies consider both previous moves and reputation for input.

Yao and Darwen [22] have earlier investigated this issue. In their study, each strategy is tagged with a reputation score, which is calculated based on payoffs received from a small, random sample of pre-games (here, we use the actual choices played by the strategy to calculate the reputation score to better reflect its cooperativeness). They used a coevolutionary learning model to show that with the addition of reputation, cooperative outcomes are possible and more likely even for the case of the IPD with more choices and shorter game durations. Note that their study is different compared with other studies such as [21]. In particular, the model constructed in [21] specifically encoded reputation (e.g., image scores) into the behavioral responses of strategies, and that reputation alone was used as the input for decision-making.

However, there are still two important issues that have not been fully addressed in [22]. First, it is not clear why with increasingly more available choices for play, the coevolutionary process is less likely to recover from the early drop to defection play starting from a random population. Although the lack of behavioral diversity was shown to be a cause [14]–[16], it is still not known why this outcome is more likely to occur when there are more available choices for play. Second, it is not known precisely why and how reputation helps with the evolution of cooperation even in cases of the IPD with more choices and also shorter game durations.

Using a coevolutionary learning model, we first show why the increasing number of available moves for play can make it more difficult for strategies to engage in highly cooperative play (i.e., full cooperation). That is, unless strategies are nice

from the beginning of the interaction and continue to be mutually cooperative thereafter, more choices discourage mutually cooperative behavior. Our experiments suggest that when there are more choices, strategies have more opportunities to exploit others by playing lower cooperation levels. For the case of the coevolutionary learning of the classical IPD where the mechanism of direct reciprocity allows the learning of cooperative behaviors, it is easy for strategies to evolve to reciprocate cooperation or to punish defection. However, with more choices, we note that strategies are less able to resolve the intention of an intermediate choice, e.g., whether it is a signal to engender further cooperation or a subtle exploitation. Strategies adapt to play lower cooperation levels that provide higher payoffs in the short-term view when they are unable to resolve the intention of opponents.

Following that, in our second study, we show why and how the addition of reputation in the IPD game makes the evolution of cooperation more likely. Our study shows that the coevolutionary learning of cooperative behaviors is possible and more likely when strategies evolved to use reputation as a mechanism to estimate behaviors of future partners and to elicit cooperation right from the start of interactions. In particular, we observe that when strategies consider both previous moves and reputation scores, our coevolutionary learning model is more likely to evolve strategies that play binary choices (e.g., full cooperation and full defection) despite the option to play other intermediate choices. Cooperative outcomes are obtained when strategies evolve to maintain high reputation scores that require highly cooperative play. Strategies are able to maintain mutual cooperation through a “discriminatory” play, i.e., fully cooperate with strategies that have good reputation (e.g., their reputation scores are similar or higher), otherwise, fully defect against strategies with bad reputation (e.g., their reputation scores are lower).

We also investigate the impact of different implementations of reputation on the coevolutionary process and why it is important for the evolution of cooperation. Experiments are carried out to determine how different methods for calculating reputation scores (based on different interpretations on how a strategy’s reputation is estimated by other strategies) affect behaviors of future interactions. In particular, we focus on the impact of accurate reputation estimation (e.g., how well a particular implementation estimates the behaviors of strategies) on the evolutionary outcome. We show that accuracy of reputation estimation depends on how memory of games from previous generations is incorporated to calculate reputation scores, and how frequently reputation scores are updated. Increasing accuracy of reputation estimation has a significant and positive impact on the evolution of cooperation.

The rest of this paper is organized as follows. Section II describes the IPD game and the coevolutionary learning model of the experiments. Section III presents the results and analysis of the experiments to explain why more choices lead to defection. Section IV details previous work in incorporating reputation into IPD, how we implement it here, and presents the computational study on why and how reputation can help with the evolution of cooperation for the IPD with more choices. Section V presents results and analysis of experiments for different imple-

	Cooperate	Defect
Cooperate	R	T
Defect	S	P

Fig. 1. The payoff matrix for the two-player IPD. The payoff listed in the lower left-hand corner of each box is assigned to the Row player choosing the move. The payoff listed in the right-hand corner of each box is assigned to the Column player choosing the move.

	Cooperate	Defect
Cooperate	4	5
Defect	0	1

Fig. 2. The payoff matrix for the two-player IPD used in this paper. $T = 5$, $R = 4$, $P = 1$, and $S = 0$.

mentations of reputation. Section VI concludes this paper with some remarks for future work.

II. METHOD OF STUDY

A. The IPD Game

1) *Classical IPD*: In the classical IPD [3], each player has two choices: cooperation and defection. The payoff a player receives depends on a payoff matrix (Fig. 1) that must satisfy the following three conditions:

- 1) $T > R$ and $P > S$ (defection always pays more);
- 2) $R > P$ (mutual cooperation beats mutual defection); and
- 3) $R > (S + T)/2$ (alternating does not pay).

There are many possible values for T , R , P , and S that satisfy the above three conditions. Here, we use $T = 5$, $R = 4$, $P = 1$, and $S = 0$ (Fig. 2).

2) *IPD With Multiple Levels of Cooperation*: The IPD with multiple levels of cooperation follows previous work in [14]–[16] and [22]. The n -choice IPD is based on a simple linear interpolation of the classic *two-choice* IPD using the following equation [16]:

$$p_A = 2.5 - 0.5c_A + 2c_B, \quad -1 \leq c_A, c_B \leq 1$$

where p_A is the payoff to player A, given that c_A and c_B are the cooperation levels of the choices that players A and B make, respectively.

In generating the payoff matrix for the n -choice IPD, the following conditions must be satisfied.

- 1) For $c_A < c'_A$ and constant c_B : $p_A(c_A, c_B) > p_A(c'_A, c_B)$.
- 2) For $c_A \leq c'_A$ and $c_B < c'_B$: $p_A(c_A, c_B) < p_A(c'_A, c'_B)$.
- 3) For $c_A < c'_A$ and $c_B < c'_B$: $p_A(c'_A, c'_B) > (1/2)(p_A(c_A, c'_B) + p_A(c'_A, c_B))$.

One can observe that the above conditions are analogous to those for the *two-choice* IPDs. For example, the first condition ensures that defection always pays more. The second condition

		PLAYER B			
		+1	$+\frac{1}{3}$	$-\frac{1}{3}$	-1
PLAYER A	+1	4	$2\frac{2}{3}$	$1\frac{1}{3}$	0
	$+\frac{1}{3}$	$4\frac{1}{3}$	3	$1\frac{2}{3}$	$\frac{1}{3}$
	$-\frac{1}{3}$	$4\frac{2}{3}$	$3\frac{1}{3}$	2	$\frac{2}{3}$
	-1	5	$3\frac{2}{3}$	$2\frac{1}{3}$	1

Fig. 3. The payoff matrix for the two-player *four-choice* IPD used in this paper. Each element of the matrix gives the payoff for Player A [14].

ensures that mutual cooperation has a higher payoff than mutual defection. The third condition ensures that alternating between cooperation and defection does not pay in comparison to just playing cooperation.

Given the payoff equation and the three conditions above, an n -choice IPD can be formulated [14]. For example, Fig. 3 shows the payoff matrix for a *four-choice* IPD. Fig. 3 also illustrates two important points. First, the payoffs in the four corners of an n -choice IPD payoff matrix are the same as those in the *two-choice* IPD. Second, any 2×2 submatrix of the $n \times n$ matrix of the n -choice IPD is itself a *two-choice* IPD.

B. Coevolutionary Learning Model

Coevolutionary learning models allow the study on the learning of game-playing strategies for the n -choice IPD. They are used in experiments of why cooperative behaviors are more difficult to evolve with more choices, and to determine the impact of reputation on evolving cooperative behaviors. We use a model based on variations of real-valued weights of a neural network [23]. In particular, we consider the self-adaptive evolutionary algorithms, e.g., evolutionary programming [24], which can be implemented as a coevolutionary procedure when fitnesses depend on interactions between members of the population [25]. The general coevolutionary learning procedure follows our earlier approach in [19]. Details of coevolutionary procedure for specific experiments are presented in the corresponding sections. Unless stated otherwise, in all the experiments, the population coevolves for 600 generations, which is sufficiently long to observe evolutionary results (e.g., persistent periods of cooperation or defection). Each experiment is repeated for 30 independent runs.

For strategy representation, we use a fixed-architecture feedforward neural networks in all the experiments for a more direct comparison with the earlier study in [22]. For the experiments with the n -choice IPD without reputation, we use the same neural network specified in [19]. As in [19] and [22], we consider deterministic and reactive memory-one strategies (i.e., response based on the moves made by the strategy and the opponent in the previous round) for simplicity and also more direct comparisons. The neural network is a multilayer perceptron with four input nodes, ten hidden nodes in the only hidden layer, and one output node. The neural network has a total of $N_w = 61$ connection weights, i.e., $[(4 \text{ input weights} + 1 \text{ bias}) \times (10 \text{ input nodes})] + [(10 \text{ connection weights} + 1 \text{ bias}) \times (1 \text{ output node})]$.

For the experiments that consider the n -choice IPD with reputation, an additional input node is added. The resulting neural network is similar to that used in [22]. Altogether, there are $N_w = 71$ connection weights, i.e., $[(5 \text{ input weights} + 1 \text{ bias}) \times (10 \text{ input nodes})] + [(10 \text{ connection weights} + 1 \text{ bias}) \times (1 \text{ output node})]$. The activation function used in the nodes is a tanh scaled to output values between $+1$ and -1 . The neural network output is rounded to the nearest value defined by the choices of the n -choice IPD game.

The five input nodes for the n -choice IPD with reputation take in the following values (for the n -choice IPD game without reputation used for comparison, only the first four input nodes are used in the neural network).

- 1) The neural network's previous choice, i.e., level of cooperation, in $[-1, +1]$.
- 2) The opponent's previous level of cooperation.
- 3) An input of $+1$ if the opponent played a lower cooperation level compared with the neural network, and 0 , otherwise.
- 4) An input of $+1$ if the neural network played a lower cooperation level compared with the opponent, and 0 , otherwise.
- 5) An input of $+1$ if the opponent has an equal or higher reputation score compared with the neural network's, and -1 , otherwise.

Note that during the reputation estimation stage, the fifth input node takes in a 0 value. In addition to the neural network, each strategy also stores two values used as pre-game inputs for a reactive memory-one IPD strategy.

III. WHY MORE CHOICES LEAD TO DEFECTION

A. Coevolutionary Learning Model for the IPD With More Choices

The coevolutionary learning model used in the experiment to investigate why more choices lead to defection outcomes in the IPD with more choices is detailed as follows.

- Step 1) Generation step $t = 0$.
Initialize $N/2$ parent strategies, P_i , $i = 1, 2, \dots, N/2$, randomly. For each strategy i :
- a) Weights and biases $[w_i(j)]$ are initialized with random values uniformly distributed in the range of $[-1, 1]$.
 - b) Each component for the self-adaptive parameter vector $[\sigma_i(j)]$ is initialized to 0.05 , e.g., a small value compared with the magnitude of the initial random connection weights.
 - c) The two pre-game inputs are initialized randomly and can take only the n -choices value uniformly distributed in that range of $[-1, 1]$.
- Step 2) Generate $N/2$ offspring, O_i , $i = 1, 2, \dots, N/2$, from $N/2$ parents. For each offspring, O_i ($[w'_i(j)]$ and $[\sigma'_i(j)]$), generated from parent, P_i ($[w_i(j)]$ and $[\sigma_i(j)]$) through a self-adaptive Gaussian mutation
- $$\sigma'_i(j) = \sigma_i(j) * \exp(\tau * N_j(0, 1)),$$
- $$i = 1, \dots, N/2; \quad j = 1, \dots, N_w$$
- $$w'_i(j) = w_i(j) + \sigma'_i(j) * N_j(0, 1),$$
- $$i = 1, \dots, N/2; \quad j = 1, \dots, N_w$$

where $N_w = 61$, $\tau = (2(N_w)^{0.5})^{-0.5} = 0.2530$, and $N_j(0, 1)$ is a Gaussian random variable (zero-mean and standard deviation of one) resampled for every j . N_w is the total number of weights, and biases for the neural network. Each of the two pre-game inputs is mutated separately. Mutation is of the form of adding the original value with the step value having an approximately standard Gaussian distribution. The new value does not wrap around, e.g., values exceeding $+1$ are changed to $+1$, while values that are lower than -1 are changed to -1 .

- Step 3) All pairs of strategies compete in the tournament, including the case where strategies play with themselves (e.g., round-robin tournament). For N strategies in a population, every strategy competes a total of N games. There are $(N^2 + N)/2$ games that are played in total.
- Step 4) Select the best $N/2$ strategies based on total payoffs of all games played. Increment generation step $t = t + 1$.
- Step 5) Steps 2 to 4 are repeated until the termination criterion (i.e., a fixed number of generation) is met.

B. Revisiting the IPD With More Choices: Evolution of Defection

Many previous studies such as [13]–[16] have shown that the evolution of defection is more likely for the IPD with more choices. Their results suggest that the mechanism of direct reciprocity is less effective in promoting the learning of cooperative behaviors through coevolution. Here, we reexamine the studies with a set of controlled experiments aimed to further investigate this issue.

We consider two experiments. The first experiment uses the 64-choice IPD, while the second experiment uses the four-choice IPD. Both experiments have a game duration of ten rounds. Note that in the first experiment, the number of choices is larger than the number used for the game duration (e.g., 64 choices compared with ten rounds), while the opposite applies in the second experiment where the number used for the game duration is larger than the number of choices. These experiments allow us to study whether the mechanism of direct reciprocity for the coevolutionary learning of IPD games with significantly more choices (e.g., the number of choices is more than the game duration) is as effective as the case of IPD games with a small number of intermediate choices (e.g., the number of choices is less than the game duration, which is the case for most studies that considered the classical IPD with two choices).

Tables I and II summarize the results of the experiments. Results show that populations are more likely to evolve to mutual defection play for the case of the 64-choice IPD game where the number of choices is larger compared with the number used for the game duration than for the case of the four-choice IPD game where the opposite is true. For the experiment that used the 64-choice IPD, 25 runs out of 30 resulted with defection (Table II). However, when there are only four choices, only nine out of 30 runs resulted with defection (Table II). The comparison of population average payoffs of individual runs between

TABLE I

COMPARISON OF RESULTS FROM THE EXPERIMENTS THAT USED THE IPD WITH FOUR CHOICES AND 64 CHOICES. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. "MEAN" INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. "STD DEV" INDICATES THE STANDARD DEVIATION. "MAX" INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. "MIN" INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE p -VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE *FOUR*-CHOICE IPD EXPERIMENT WITH THAT OF THE 64-CHOICE IPD EXPERIMENT

Choices	Mean $\pm$ Std Dev	Max	Min	p -value
4	2.27 $\pm$ 1.13	3.98	1.01	
64	1.37 $\pm$ 0.84	3.87	1.00	0.0010 [†]

[†] The p -value indicates a statistically significant result for a 0.05 level of significance that is considered.

TABLE II

COMPARISON OF RESULTS FROM THE EXPERIMENTS THAT USED THE IPD WITH FOUR CHOICES AND 64 CHOICES. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. " $No \leq 1.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFECTION PLAY. " $1.5 < No < 3.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. " $No \geq 3.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Choices	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
4	9	13	8
64	25	4	1

the two experiment sets shows a statistical significant difference in the outcomes of the two experiments (Table I).

Table III summarizes statistics that describe behavioral responses of final evolved and most fit strategies in the populations. From the table, we observed that there are more runs where evolved strategies are nasty (e.g., play lower cooperation levels at the start of the game) for the case of the 64-choice IPD compared with the case of the *four*-choice IPD (e.g., 26 runs in the 64-choice IPD experiment compared with 19 runs in the *four*-choice IPD experiment). We also observed that there are more runs where evolved strategies respond with lower cooperation levels if previous moves indicated mutual cooperation play for the experiment with more choices (e.g., 24 runs in the 64-choice IPD experiment compared with 13 runs in the *four*-choice IPD experiment). This indicates that highly cooperative play is more difficult to sustain since strategies are not willing to continuously engage in mutual cooperation. Finally, Table III also shows that final evolved strategies are not naive (e.g., respond with higher cooperation levels if previous moves indicated mutual defection play) in all the runs for both experiments. These strategies respond with defective choices if in the previous rounds, mutual defection occurred in the previous round.

Closer inspection of final evolved and most fit strategies in the populations of most runs that evolved to defection for the 64-choice IPD experiment revealed that strategies mostly started the game with low cooperation levels and continued to respond with lower cooperation levels (e.g., Fig. 4). We

also observed in most runs with defection outcomes that the populations consisted of similar strategies.

Results from the two controlled experiments suggest that the mechanism of direct reciprocity is less effective in promoting the evolution of cooperative behaviors when the number of choices increases (which is more apparent when the number of choices is more than the game duration). We have conducted two more controlled experiments to illustrate this point further. In particular, we change the equation from $p_A = 2.5 - 0.5c_A + 2c_B$ to $p_A = 2.5 - 0.25c_A + 2.25c_B$. This results with IPD payoffs that reduce the penalty given for playing higher cooperation levels (e.g., for the classical IPD, $T = 5$, $R = 4.5$, $P = 0.5$, and $S = 0$), and should encourage the evolution of cooperation.

Results from these two experiments are summarized in Tables IV–VI. Comparing results from these two new experiments (Tables IV–VI) that use payoffs that penalize higher cooperation levels less than the previous two experiments (Tables I–III), we observed that cooperation outcomes are more likely for the new experiments regardless of the number of choices in the IPD. However, even with payoffs that encourage more cooperation play, the increase in the number of choices leads to the evolution of defection. Tables V and VI show that strategies tend to play lower cooperation levels when the number of choices increase from four to 64, and that comparison of the outcomes from the two experiments shows a statistical significant difference (Table IV).

C. More Choices Create More Opportunities to Exploit Opponents

Earlier, our experiments show that populations are more likely to evolve to play mutual defection when there are more choices available for play in the IPD game. Here, we investigate why more choices lead to the evolution of defection. In addition to investigating individual runs, we also collect statistical evidence for general observations on how strategies explore or sample choices during game-play in experiments that used the 64-choice IPD and the *four*-choice IPD. We use the *entropy* formula to calculate the distribution of sampled choices of play of a strategy's behavioral map. For example, there are $n_t = n^2$ elements in the behavioral response map of a memory-one reactive strategy for an n -choice IPD. The sampling distribution can be calculated as follows:

$$-\sum_l \frac{m_l}{M} * \log_2 \frac{m_l}{M} \quad (1)$$

where m_l/M is the frequency that the element l of the strategy map is sampled for a total of $M = \sum_l m_l$ responses. Equation (1) is applied to the most fit strategy in each generation for simplicity in monitoring how strategies explore choices during the evolutionary process.

With the entropy value unnormalized, higher values indicate more alternative responses to different combinations of previous moves that are sampled frequently by strategies. Furthermore, given that evolved strategies are unlikely to play

TABLE III
COMPARISON OF FINAL EVOLVED AND MOST FIT STRATEGY PROPERTIES IN THE EXPERIMENTS THAT USED THE IPD WITH FOUR AND 64 CHOICES. THE “FIRST MOVE” COLUMN COLLECTS THE NUMBER OF RUNS, OUT OF 30, WHERE THE STRATEGY STARTS A GAME WITH A “NICE” MOVE (E.G., COOPERATION LEVELS ≥ 0), OR A NASTY MOVE (E.G., COOPERATION LEVELS < 0). THE “RESPONSE TO MUTUAL COOPERATION” COLUMN COLLECTS THE NUMBER OF RUNS, OUT OF 30, WHERE THE STRATEGY RESPONDS TO PREVIOUS MUTUAL COOPERATION PLAY A “COOPERATE” MOVE (E.G., COOPERATION LEVELS ≥ 0), OR A “DEFECT” MOVE (E.G., COOPERATION LEVELS < 0). SIMILARLY, THE “RESPONSE TO MUTUAL DEFECTION” COLUMN COLLECTS STATISTICS FOR STRATEGY RESPONSE TO MUTUAL DEFECTION PLAY

Choices	First Move		Response to Mutual Cooperation		Response to Mutual Defection	
	Nice	Nasty	Cooperate	Defect	Defect	Cooperate
4	11	19	17	13	30	0
64	4	26	6	24	30	0

Fig. 4. Most fit strategy's behavioral response that evolved to defection play in the 64-choice IPD experiment for a typical run with defection outcome. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 0.

TABLE IV
COMPARISON OF RESULTS FROM THE EXPERIMENTS THAT USED THE IPD WITH FOUR CHOICES AND 64 CHOICES WITH PAYOFFS THAT PENALIZE HIGHER COOPERATION LEVELS LESS. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “MEAN” INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. “STD DEV” INDICATES THE STANDARD DEVIATION. “MAX” INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. “MIN” INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE p -VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE *FOUR*-CHOICE IPD EXPERIMENT WITH THAT OF THE *64*-CHOICE IPD EXPERIMENT

Choices	Mean $\pm$ Std Dev	Max	Min	p -value
4	2.79 ± 1.57	4.49	0.50	
64	1.72 ± 1.60	4.50	0.50	0.0076 [†]

[†] The p -value indicates a statistically significant result for a 0.05 level of significance that is considered.

a single-choice response (e.g., “all cooperate” in a cooperative population), higher entropy values would indicate that strategies play more alternative choices. In a comparison between strategies of IPD games with different number of choices, the emphasis is on how much more alternative choices are played rather than the proportion of available choices that are played (where entropy needs to be normalized). As such, unnormalized entropy provide a quantitative measure of how strategies “invest” interaction opportunities (which are similar for IPD

TABLE V
COMPARISON OF RESULTS FROM THE EXPERIMENTS THAT USED THE IPD WITH FOUR CHOICES AND 64 CHOICES WITH PAYOFFS THAT PENALIZE HIGHER COOPERATION LEVELS LESS. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “ $No \leq 1.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFECTION PLAY. “ $1.5 < No < 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. “ $No \geq 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Choices	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
4	8	9	13
64	19	4	7

games with different number of choices) in playing alternative choices.

We first compare the average entropy values over 30 runs for the two experiments using IPD games with a different number of choices to provide some general quantitative information as to how choices are played by evolving strategies. We note from Fig. 5 that the figure plots the average entropy values over 30 runs for 600 generations for the two experiments where two observations can be made. First, the average entropy values are higher for the 64-choice IPD experiment compared with the *four*-choice IPD experiment. Second, and more interestingly, there is a very rapid decrease in the average entropy values for the 64-choice IPD experiment compared with the *four*-choice IPD experiment.

We further investigate why there is such a very rapid decrease in average entropy value for the 64-choice IPD experiment. In particular, closer inspection of individual runs reveals that the populations evolved to play lower cooperation levels more frequently after starting from initial random populations. The change from an initial, well-spread distribution of sampling frequency of choices played to a narrower distribution during the early period of the evolutionary process (e.g., Fig. 6, which shows four sample runs) explains the rapid decrease of the average entropy value in Fig. 5 for the 64-choice IPD experiment.

The change in play towards lower cooperation levels is observed in almost all runs where populations evolved to mutual defection for the 64-choice IPD experiment (Table II). However, for the experiment with the *four*-choice IPD, one can observe that in addition to individuals runs where the population evolved to mutual defection, there are runs where populations

TABLE VI
COMPARISON OF MOST FIT, FINAL EVOLVED STRATEGY PROPERTIES IN THE EXPERIMENTS THAT USED THE IPD WITH FOUR AND 64 CHOICES WITH PAYOFFS THAT PENALIZE HIGHER COOPERATION LEVELS LESS. THE “FIRST MOVE” COLUMN COLLECTS THE NUMBER OF RUNS, OUT OF 30, WHERE THE STRATEGY STARTS A GAME WITH A “NICE” MOVE (E.G., COOPERATION LEVELS ≥ 0), OR A NASTY MOVE (E.G., COOPERATION LEVELS < 0). THE “RESPONSE TO MUTUAL COOPERATION” COLUMN COLLECTS THE NUMBER OF RUNS, OUT OF 30, WHERE THE STRATEGY RESPONDS TO PREVIOUS MUTUAL COOPERATION PLAY A “COOPERATE” MOVE (E.G., COOPERATION LEVELS ≥ 0), OR A “DEFECT” MOVE (E.G., COOPERATION LEVELS < 0). SIMILARLY, THE “RESPONSE TO MUTUAL DEFECTION” COLUMN COLLECTS STATISTICS FOR STRATEGY RESPONSE TO MUTUAL DEFECTION PLAY

	First Move		Response to Mutual Cooperation		Response to Mutual Defection	
Choices	Nice	Nasty	Cooperate	Defect	Defect	Cooperate
4	18	12	21	9	25	5
64	11	19	11	19	29	1

Fig. 5. Comparison of the average entropy values over 30 runs for 600 generations between the IPD with four choices and 64 choices. The average entropy values over 30 runs taken at the final generation are 0.68 and 1.34 for the IPD with four and 64 choices, respectively.

also evolved to play intermediate choices and mutual cooperation (Fig. 7).

With more choices, one would expect equal opportunities for encouraging cooperation and for exploiting partners. That is, one would expect that there are no large changes to the proportion of mutual cooperation and mutual defection outcomes from the evolutionary process. However, as with the other past studies, our experiments show that this is not true. Results always show that mutual defection outcomes are always more likely compared with mutual cooperation outcomes when there are more choices to play.

We first note that in the IPD, payoffs are higher for exploiting a cooperative partner than mutual cooperation in the short-term (e.g., for that particular round itself). With two choices and repeated interactions, the mechanism of direct reciprocity allows for the learning of cooperative behaviors as strategies can evolve to reciprocate cooperation and to punish defectors, which would discourage evolution to mutual defection due to lower average payoffs.

However, with more choices, it is more difficult to evolve strategies to reciprocate cooperation and to punish defectors (to

avoid being exploited) as it may not be entirely clear what a particular intermediate choice will signal. That is, strategies are less able to resolve an intermediate choice as a signal to engender further cooperation or a subtle exploitation. This is despite having repeated interactions because even a moderate increase in the number of choices leads to a substantially larger increase in the possible sequences of choices that makes resolving the intention of choices played earlier much more challenging.

Given that the coevolutionary process is driven by strategic interactions defined by the average payoff of all IPD games played, it would appear that there are more incentives for strategies to adapt their responses to further exploit their partners rather than to engender cooperation in the short-term when they are unable to resolve the intention of opponents. These observations would be more apparent when the number of choices is higher than the number used for the game duration. An interpretation to these observations is that when there are more choices, there is effectively more opportunities to exploit partners. The mechanism of direct reciprocity becomes less effective in promoting the learning of cooperative behaviors as strategies learn to play lower cooperation levels.

D. The Impact of Game Duration

Further experiments that use the 64-choice IPD, but with longer game durations, show that the number of runs where the population evolved to play mutual defection are reduced (Tables VII and VIII). For example, the number of runs that evolved to defection reduced from 25 to 15 after increasing game duration from 10 to 30, although there is no further decrease when game duration is further increase to 150 (Table VIII). Comparisons of the population average payoffs of individual runs between the original experiment with a game duration of ten with the experiments with game durations of 30 and 150 show that there are statistically significant differences (Table VII).

We note that although increasing the game duration reduces the number of runs with defection outcomes, there is no positive and significant impact to the number of runs with cooperation outcomes. Instead, there are more runs where the outcomes are characterized by either alternative or intermediate choices of plays (Table VIII). This result is not unexpected given that

Fig. 6. Sampling distribution of choices of the populations of the 64-choice IPD experiment. In general, most runs converged to defection play within 40 generations, e.g., (a)–(d). Choices are represented from 0 (full defection) to 63 (full cooperation). A higher vertical line for a particular choice and a particular generation indicates a higher sampling frequency for the choice. (a) Run # 1. (b) Run # 2. (c) Run # 3. (d) Run # 4.

the IPD game with 64 choices leads to many more possible sequences of choices. For example, Fig. 8 shows that higher average entropy values for experiments with longer game durations suggest that strategies evolve to play other choices.

E. Is Cooperative Play Still Possible?

At this point, we would like to find out whether it is possible for strategies to engage in highly cooperative play through some other mechanisms for the IPD with more choices. Instead of descending to lower cooperation play and staying there (e.g., Fig. 6), we investigate if strategies can evolve to climb their way back to higher cooperation play. For example, with the coevolutionary learning of the *two-choice* IPD, “tit-for-tat” type of strategies have the properties to bring a population playing defection back to cooperation if there is sufficient payoff gains from cooperating with strategies that cooperate compared with the loss of payoffs from strategies that defect (e.g., being exploited in the first round by “all defect”).

Here, we consider the “Raise-the-Stakes” mechanism proposed by Roberts and Sheratt [26] when interactions involve

more available choices. In their study, it was shown that strategies that increase their cooperation levels over the duration of their encounters can encourage cooperative play (including self-play where both players raise their cooperation levels until they reach mutual cooperation play). Here, we attempt to find evidence of strategies employing this mechanism since the representation that we use can adopt “Raise-the-Stakes” strategies by playing increasingly higher cooperation levels compared with cooperation levels played in the previous round.

According to the study [26], such strategies will have to rapidly increase its cooperativeness after starting from a relatively low cooperation level play. However, a “Raise-the-Stakes” strategy can win more, or lose more, depending on how it balances raising the cooperation levels in the following moves with similar strategies to quickly obtain the benefit of a mutual cooperation play without being vulnerable to exploitation by others that stop raising their cooperation levels midway.

For the 64-choice IPD experiment, we observed that there was only one out of 30 runs where the population evolved to mutual

Fig. 7. Sampling distribution of choices of the populations in the *four-choice* IPD experiment. Different changes to the sampling distribution can be observed, e.g., (a) changes to intermediate choices, (b)-(c) mutual cooperation, and (d) mutual defection. Choices are represented from 0 (full defection) to 3 (full cooperation). A higher vertical line for a particular choice and a particular generation indicates a higher sampling frequency for the choice. (a) Run # 1. (b) Run # 2. (c) Run # 3. (d) Run # 4.

TABLE VII

COMPARISON OF RESULTS FROM THE EXPERIMENTS THAT USED THE 64-CHOICE IPD WITH LONGER GAME DURATIONS. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “MEAN” INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. “STD DEV” INDICATES THE STANDARD DEVIATION. “MAX” INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. “MIN” INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE *p*-VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE EXPERIMENT WITH A GAME DURATION OF TEN ROUNDS AND THAT OF EXPERIMENTS WITH OTHER GAME DURATIONS

Choices (RoundNo)	Mean $\pm$ Std Dev	Max	Min	<i>p</i> -value
64 (10)	1.37 $\pm$ 0.84	3.87	1.00	
64 (20)	1.63 $\pm$ 0.87	3.52	1.00	0.2537
64 (30)	1.89 $\pm$ 0.94	3.50	1.00	0.0197 [†]
64 (150)	1.80 $\pm$ 0.91	3.94	1.00	0.0035 [†]

[†] The *p*-value indicates a statistically significant result for a 0.05 level of significance that is considered.

TABLE VIII

COMPARISON OF RESULTS FROM THE EXPERIMENTS THAT USED THE 64-CHOICE IPD WITH LONGER GAME DURATIONS. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “ $No \leq 1.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFECTION PLAY. “ $1.5 < No < 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. “ $No \geq 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Choices (RoundNo)	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
64 (10)	25	4	1
64 (20)	19	10	1
64 (30)	15	14	1
64 (150)	15	13	2

cooperative play at the end of the run after initial lower cooperation play. In that particular run, the final evolved and most

fit strategy used a form of “Raise-the-Stakes” strategy (Fig. 9). We also observed that the population at the end of this run consisted mostly of strategies similar to this strategy. Although the behavioral map shown in Fig. 9 indicates three different regions, the broad region of full cooperation responses on the right is the

Fig. 8. Comparison of the average entropy values over 30 runs over 600 generations between the 64-choice IPD with various game durations. The average entropy values over 30 runs taken at the final generation are 1.34, 1.96, 2.16, and 2.02 for the 64-choice IPD with game duration of 10, 20, 30, and 150, respectively.

Fig. 9. Most fit evolved strategy's behavioral response that evolved to highly cooperative play in the 64-choice IPD experiment. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 34.

most important. This region represents a sort of tolerance of the strategy to alternative choices since it continues to respond with full cooperation regardless of the opponent's previous move that is lower in cooperation levels.

However, closer inspections on the evolution of this particular run showed that cooperative play did not emerge from "Raise-the-Stakes" strategies invading the population. Consider Fig. 10, which plots the three most indicative properties (e.g., the first move, the response to mutual cooperation, and the response to mutual defection) for the evolved and most fit strategies throughout the evolutionary run. The figure shows that the transition from low to higher cooperation plays first began with idiosyncratic behaviors where strategies started their games playing nice, but which were quickly followed with low cooperation play. For example, Fig. 10 shows that evolution of full cooperation play started with the full cooperation play for the first move at around generation 256 (top graph). However, it was only nine generations later (middle graph) that full cooperation play was evolved for the response to previous mutual cooperation play. We also observed that a

Fig. 10. Properties of most fit evolved strategies in the run that evolved to cooperation. The top graph plots the choices played for the first move. The middle graph plots the choices played if previous moves are mutual cooperation (full cooperation play of both the strategy and the opponent). The lower graph plots the choices played if previous moves are mutual defection (full defection play of both the strategy and the opponent).

Fig. 11. Sampling distribution of choices of the population in the 64-choice IPD experiments throughout the evolutionary process of the only run that converged to cooperation. Choices are represented from 0 (full defection) to 63 (full cooperation). A higher vertical line for a particular choice and a particular generation indicates a higher sampling frequency for the choice. The graph plots the sampling distributions from generations 250 to 280 to show the transition from mostly defective play to higher cooperative play.

large number of offspring strategies playing idiosyncratically around generation 256.

After that, strategies evolved to a "tit-for-tat" strategy that started the game with a full cooperation play, and then reciprocated the opponent's cooperation levels. It was only around 60 generations later that "Raise-the-Stakes" strategies that were similar to that shown by Fig. 9 appeared in the population. The transition from lower to higher cooperative play can also be seen in Fig. 11, which plots the sampling distribution of choices played by the population.

Although a form of "Raise-the-Stakes" strategy was evolved towards the end the run (Fig. 9), cooperation did not emerge

from these strategies. Rather, the transition to higher cooperative play was the result of the population introducing a large number of idiosyncratic strategies that started by playing nice, but did not reciprocate cooperation. These strategies then became stepping stones for evolution to produce cooperative strategies, which then further evolved to strategies that play “Raise-the-Stakes.”

Our results indicate that it is unlikely for a coevolving population to climb its way back to mutual cooperation from mutual defection using mechanisms such as “Raise-the-Stakes.” There are two possible reasons for this observation. First, there is a delicate balance that needs to be met in how fast a strategy can raise the level of cooperation without making itself vulnerable to exploitation [26]. Second, the strategy representation that is used here allows for complex nonlinear responses that can exploit some form of “Raise-the-Stakes” strategy.

Instead, the coevolutionary learning of cooperative behaviors from mutual defection play seems to depend first on strategies engaging in highly cooperative play right from the start of the interaction. Strategies later evolved to cooperative play through a form of “Raise-the-Stakes” strategy. There are other studies that also reported similar observations. For example, Wahl and Nowak [17] also concluded that offering high cooperation levels at the start of the interaction is important for the evolution of cooperative plays because it avoids the need for “building up trust” by raising the stakes.

IV. INCORPORATING REPUTATION INTO THE ITERATED PRISONER’S DILEMMA (IPD)

A. Motivation

For the coevolutionary learning of strategies in the classical IPD game, strategies can learn cooperative behaviors from the mechanism of direct reciprocity. However, for the IPD with more choices, the evolution of cooperation is less likely, suggesting a failure in the mechanism in promoting the learning of cooperative behaviors. Given that complex human interactions are not limited to direct interactions [20], the coevolutionary learning framework for IPD may require further extension to model complex interactions. With the extended framework, strategies may learn cooperative behaviors without relying entirely on the mechanism of direct reciprocity.

One mechanism that allows the evolution of cooperation, but does not require direct interactions, is reputation [21]. Reputation is argued and shown to be the mechanism for indirect reciprocity to occur [21], [27]. Indirect reciprocity is understood to occur when cooperation between current partners depends on their prior behaviors to others. Unlike the case of direct reciprocity where cooperation depends on repeated encounters between individuals (e.g., cooperative behaviors in interactions modeled as IPD games), in the case of indirect reciprocity, an individual receives cooperation from third parties due to the individual’s cooperative behaviors to others. Here, we explore the idea of incorporating reputation in the IPD game. Indeed, it has been mentioned earlier that in complex interactions such as those found in human societies, both mechanisms of direct and indirect reciprocity exist together, with the mechanism of

indirect reciprocity occurring as a result of direct reciprocity between other individuals [20].

Yao and Darwen [22] first studied reputation in the IPD game with more choices. Although results from their experiments indicate that evolution of cooperation is more likely with reputation, even for cases with short game durations, their study did not explain why and how reputation helps with evolving cooperative behaviors. Here, we will analyze the reasons behind reputation in promoting cooperation in the coevolutionary learning of IPD strategies.

B. Related Work

Nowak and Sigmund [21] studied how cooperative behaviors can be evolved through indirect reciprocity using image scoring strategies. The two-player interaction is modeled as a single-round donation game. Strategies are randomly selected to be donors and recipients. When strategies are selected to be donors, they are given two choices: cooperate (help) and defect (not help). For any donor-recipient pair, the donor pays a cost of c if it cooperates, while the recipient receives the benefit b , with $b > c$. If the donor defects, both donor and recipient receives zero payoff.

Each strategy is given a value for reputation that is known to the strategy and its opponents. There are two ways for calculating reputation: image scores [21] and standing [27]. For both methods, an individual’s reputation score increases if the donor cooperates. The two methods differ in terms of how the reputation score is decreased when the donor defects. For image score, the reputation score decreases whenever the donor defects [21]. For standing, the reputation score only decreases when the defection is unjustified (e.g., the recipient has a good reputation score). Otherwise, the reputation score remains unchanged [27].

A strategy’s behavioral response depends on the reputation score and some additional parameters used for decision-making (e.g., the strategy representation is a set of values that includes its reputation score and parameters for thresholding). Previous moves are not used in the decision-making process. Depending on how strategies use information from reputation scores, there are three broad classes of strategies [27]. First, a strategy can consider its own reputation score only for decision-making (e.g., “offer help when own reputation score is less than h ”). Second, a strategy can consider the recipient’s reputation score for decision-making (e.g., “offer help when recipient’s reputation score is at least k ”). Third, a strategy can consider both its own and the recipient’s reputation scores (e.g., “offer help when own reputation score is less than h and when recipient’s reputation score is at least k ”) [27].

Our approach, and that of [22], differs from previous studies [21], [27] on three fundamental aspects. First, in terms of the interaction, we use a different game compared with [21] and [27]. Here, we consider the IPD game with multiple levels of cooperation. In [21] and [27], the game is a single-round donation game with only two extreme choices. Second, our approach is different in terms of a strategy’s behavioral response. We consider a strategy whose responses depend on both previous moves and reputation. In [21] and [27], only reputation is used as inputs to a strategy’s response. Third, we use a coevolutionary learning model whereby strategies can adapt their behaviors over a range

specified by the representation. Studies in [21] and [27] did not consider a process of adaptation, e.g., strategies are predetermined and their replacement depends on the proportions of existing strategies in the population.

C. Implementation of Reputation

Following the implementation for incorporating reputation in the IPD with more choices by Yao and Darwen [22], we consider a binary reputation case, i.e., good and bad reputation. How a strategy decides whether its opponent has good or bad reputation is by comparing their reputation scores. If the opponent has an equal or higher reputation score compared with the strategy's, then the strategy considers the opponent as having a good reputation. Otherwise, the strategy considers the opponent as having a bad reputation.

The implementation of reputation through comparison of reputation scores (e.g., relative rather than absolute values for reputation) follows the motivation of earlier studies [21], [27] on the mechanism of indirect reciprocity as a means to promote cooperation. In particular, these studies argued for the mechanism of indirect reciprocity as a sort of moral or cultural system that is imposed on the interacting individuals to encourage cooperation rather than being an emergent property from evolutionary processes [21], [27]. For our implementation here, the use of relative values of reputation scores provides a well-defined and efficient means for strategies to identify partners that have earlier paid the cost (e.g., leaving themselves open to exploitation) through highly cooperative plays to obtain a large reputation score that is not possible if the implementation uses absolute values of reputation scores.

Since our strategy considers both previous moves and reputation for inputs, the strategy's reputation score has to be determined first. How a strategy's reputation is estimated depends on how reputation is interpreted to reflect a strategy's possible behaviors to future opponents in light of choices it made with past opponents. Here, we focus on the issue of the accuracy of the procedure in estimating the reputation that reflects the strategy's behavior. The accuracy of reputation estimation may be related to how memory of games used to calculate reputation scores and how frequent reputation scores are updated in providing feedback to other strategies in the population.

We first examine a two-stage procedure that is first proposed in [22] for the IPD with more choices and reputation. In the first stage, which we named the reputation estimation stage, a strategy's reputation score is calculated based on choices the strategy played in some random n -choice IPD games in the current generation. After obtaining reputation scores, strategies compete in the second stage of the n -choice IPD with reputation.

It should be noted that the approach here (and also [22]) may not be realistic due to two assumptions made on how reputation is estimated. First, it is assumed that there is no memory of games played in previous generations, i.e., a strategy's reputation score is always calculated anew every generation, and choices it made in the past has no impact on its current reputation. Second, it is assumed that a strategy reputation is static for the duration of its interactions with other strategies in the population until the next round of update, which occurs in the

next generation. However, the approach does simplify analyzing the dynamics of the game by separating the procedure for estimating reputation from the strategy interactions that involved both previous moves and reputation scores.

The procedure for calculating the reputation score is described in more detail here. In particular, consider that in a population of N strategies, there are $(N^2 + N)/2$ unique pairwise interactions. Each of these pairwise interactions has a fixed probability of being sampled (without replacement) for the reputation estimation stage. Note that some strategies may play more games than others, in which case, normalization based on the number of games played is required. The reputation score for each strategy is calculated as follows.

- Each choice is given a weight. Weights are assigned in the same way as in calculating the payoffs (e.g., for four choices, $-1, -1/3, +1/3, +1$).
- The number of times a choice is played by the strategy during interactions is recorded.
- Reputation score for each strategy is obtained by taking the sum of choices played, normalized over the total number of plays the strategy makes in all the random encounters for this stage.

A formal way to describe how reputation is estimated for each strategy is to consider the distribution of choices the strategy plays in all the random encounters during the reputation estimation stage. If $f(x)$ gives the frequency distribution of the n choices, $x = \{x_1, x_2, \dots, x_n\}$, played by a strategy during the reputation estimation stage, and that the value of x specifies the weight for each choice, then the reputation score R is given by

$$R = \sum_{i=1}^n x_i * f(x_i). \quad (2)$$

The most important difference between our implementation and that of [22] is how the reputation score is calculated. In [22], a strategy's reputation score is calculated based on the payoffs received. Here, a strategy's reputation score is calculated based on the choices that a strategy plays. We believe this approach to be more accurate compared with the approach used in [22] because it better reflects the actual cooperativeness of strategies. In [22], the use of average payoffs as an indication of cooperativeness can be misleading, e.g., higher average payoff can be a result of mutual cooperation or that one strategy is exploiting the other.

D. Coevolutionary Model for the IPD With More Choices and Reputation

In the case of experiments whereby strategies also consider reputation in addition to previous moves, we used the following model.

Step 1) Generation step $t = 0$.

Initialize $N/2$ parent strategies, P_i , $i = 1, 2, \dots, N/2$, randomly. For each strategy i :

- a) Weights and biases $[w_i(j)]$ are initialized with random values uniformly distributed in the range of $[-1, 1]$.
- b) Each component for the self-adaptive parameter vector $[\sigma_i(j)]$ is initialized to 0.05, e.g., a

small value compared with the magnitude of the initial random connection weights.

- c) The two pre-game inputs are initialized randomly and can take only the n -choices value uniformly distributed in that range of $[-1, 1]$.

Step 2) Generate $N/2$ offspring, O_i , $i = 1, 2, \dots, N/2$, from $N/2$ parents. For each offspring, O_i ($[w'_i(j)]$ and $[\sigma'_i(j)]$), generated from parent, P_i ($[w_i(j)]$ and $[\sigma_i(j)]$) through self-adaptation

$$\begin{aligned}\sigma'_i(j) &= \sigma_i(j) * \exp(\tau * N_j(0, 1)), \\ i &= 1, \dots, N/2; \quad j = 1, \dots, N_w \\ w'_i(j) &= w_i(j) + \sigma'_i(j) * N_j(0, 1), \\ i &= 1, \dots, N/2; \quad j = 1, \dots, N_w\end{aligned}$$

where $N_w = 71$, $\tau = (2(N_w)^{0.5})^{-0.5} = 0.2436$, and $N_j(0, 1)$ is a Gaussian random variable (zero-mean and standard deviation of one) resampled for every j . N_w is the total number of weights, and biases for the neural network. Each of the two pre-game inputs is mutated separately. Mutation is of the form of adding the original value with the step value having an approximately standard Gaussian distribution. The new value does not wrap around, e.g., values exceeding $+1$ are changed to $+1$, while values that are lower than -1 are changed to -1 .

Step 3) Reputation score, R_{current} , for each strategy in the population (parent and offspring), is initialized to 0. In addition, each strategy's frequency distribution of choices played, $f(x)$, $x = \{x_1, x_2, \dots, x_n\}$, is initialized to 0 for all x .

Step 4) Strategies compete with each other in a two-stage game.

- a) Stage 1 (Reputation Estimation):
 - i) For every unique $(N^2 + N)/2$ pairwise interactions (games) between N strategies ($N/2$ offspring + $N/2$ parents), there is a fixed probability, p_r , a pair is sampled (without replacement) for this stage.
 - ii) Interactions are in the form of n -choice IPD game.
 - iii) For each strategy, the reputation score is calculated after the strategy has finished competing with all the randomly selected opponents, which is given as follows:

$$R_{\text{current}} = \frac{\text{Sum of all choices made}}{\text{Total number of moves made}} = \sum_{i=1}^n x_i * f(x_i) \quad (3)$$

which is obtained from (2). Strategies that have not obtained reputation scores are treated as having 0 for their reputation scores.

- iv) Each strategy is also given IPD payoffs from the interactions.
- b) Stage 2 (n -choice IPD with reputation):

- i) For the remaining pairs not sampled for Stage 1, interactions are in the form of n -choice IPD with reputation.
- ii) Each strategy is given IPD payoffs from the games.

Step 5) Select the best $N/2$ strategies based on combined payoffs from Stage 1 and 2. Increment generation step $t = t + 1$.

Step 6) Steps 2 to 5 are repeated until the termination criterion (i.e., a fixed number of generation) is met.

Comparing our implementation here with that used in [22], there are several noticeable differences. In terms of how the reputation score of a strategy is calculated, our approach uses choices that a strategy plays, while in [22], the reputation score is calculated based on payoffs received from playing a fixed number of random games. In addition, our approach also differs in how random games are selected. Here, we use a fixed probability to determine whether two strategies will participate in the reputation estimation stage. In [22], each strategy plays a fixed number of random games, with the implementation requiring random shuffling of strategies' positions in the population. In terms of coevolutionary learning models, we evolved neural network weights directly, while in [22], neural networks were evolved using genetic algorithms (e.g., variation operators operate on neural networks coded as binary strings of 0s and 1s).

It should be noted that selection of strategies are based on combined payoffs of games from both stages (reputation estimation and n -choice IPD with reputation) of the two-stage procedure described earlier. In general, the probability for selecting random games for the reputation estimation stage, p_r , should be small and less than 0.5. This procedure is consistent with our motivation of explicitly modeling both mechanisms of direct and indirect reciprocity in the interaction, and how the mechanism of indirect reciprocity occurs as a result of direct reciprocity between other individuals [20]. We will discuss the implication of combining payoffs of games from the two stages later.

E. What Are the Effects of Reputation?

1) *Reputation Leads to Cooperative Play:* Earlier results by Yao and Darwen [22] suggested that reputation can help with the evolution of cooperation for the IPD games with more choices, and also games with shorter durations. Here, we re-examine the experiments conducted in [22], whereby strategies consider both previous moves and reputation scores at the same time. We want to find out why and how reputation can help with the evolution of cooperation in IPD games with more choices and even for the case of short game durations. For more direct comparisons between experiments with and without reputation, we use the same 64-choice IPD with ten rounds of game duration.

Tables IX and X summarizes the results of the comparison. In general, our results show that when reputation was included, evolution of cooperation was more likely to occur. For example, we observed that for the experiment with reputation (and where a small value of $p_r = 0.05$ is used), there are less outcomes with

TABLE IX

COMPARISON OF RESULTS OF THE EXPERIMENTS THAT CONSIDER THE 64-CHOICE IPD GAME WITH AND WITHOUT REPUTATION. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. "MEAN" INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. "STD DEV" INDICATES THE STANDARD DEVIATION. "MAX" INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. "MIN" INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE p -VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE EXPERIMENT WITHOUT REPUTATION AND THAT OF EXPERIMENTS WITH REPUTATION AND VARIOUS p_r VALUES

Experiment	Mean $\pm$ Std Dev	Max	Min	p -value
<i>NoRep</i>	1.37 ± 0.84	3.87	1.00	
<i>Rep</i> , $p_r = 0.05$	2.20 ± 1.31	4.00	1.00	0.0751
<i>Rep</i> , $p_r = 0.15$	2.93 ± 1.40	4.00	1.00	$< 0.0001^\dagger$
<i>Rep</i> , $p_r = 0.25$	2.60 ± 1.51	4.00	1.00	$0.0038^\dagger$
<i>Rep</i> , $p_r = 0.50$	2.08 ± 1.45	4.00	1.00	0.2660

† The p -value indicates a statistically significant result for a 0.05 level of significance that is considered.

TABLE X

COMPARISON OF RESULTS OF THE EXPERIMENTS THAT CONSIDER THE 64-CHOICE IPD GAME WITH AND WITHOUT REPUTATION. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. " $No \leq 1.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFLECTION PLAY. " $1.5 < No < 3.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. " $No \geq 3.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Experiment	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
<i>NoRep</i>	25	4	1
<i>Rep</i> , $p_r = 0.05$	16	6	8
<i>Rep</i> , $p_r = 0.15$	10	0	20
<i>Rep</i> , $p_r = 0.25$	14	0	16
<i>Rep</i> , $p_r = 0.50$	19	0	11

mutual defection play and more outcomes with mutual cooperation play compared with the experiment without reputation (Table X).

However, there appears to be a tradeoff in the setting of p_r . An initial increase of p_r from 0.05 to 0.15 leads to more cooperative outcomes although further increase of p_r does not lead to further increase in the number of cooperative outcomes. Regardless, we observed comparable results between the outcomes of the coevolution of strategies for the 64-choice IPD with reputation in Table X and that of the *four*-choice IPD without reputation in Table II. This suggests that the impact of more choices in evolving mutual defection play is mitigated when reputation is incorporated.

2) *Why Reputation Helps: Strategy Responding With Binary Choices:* We start with the analysis of the sampling distribution of strategy's behavioral maps. We compare results for IPD games with and without reputation. For example, we plot the average entropy values throughout the evolution for both experiments with reputation (where $p_r = 0.05$ setting is used) and without reputation in Fig. 12. Similar observations are made for most cases of individual plots for the evolutionary runs that included reputation and that of other p_r settings. Results shown by Fig. 12 suggest that with the inclusion of reputation, strategies played less alternative choices by sampling small portions of their behavioral maps.

Fig. 12. Comparison of the average entropy values over 30 runs for 600 generations between the IPD games that consider 64 choices with and without reputation. The average entropy values over 30 runs taken at the final generation are 1.34 and 0.40 for the 64-choice IPD without reputation and with reputation, respectively.

Fig. 13. Sampling distribution of choices of the population in the case of the IPD with 64 choices and reputation throughout the evolutionary process of a typical run that converged to cooperation. Choices are represented from 0 (full defection) to 63 (full cooperation). A higher vertical line for a particular choice and a particular generation indicates a higher sampling frequency for the choice. The graph plots the sampling distributions from generations 1 to 50 to show the transition from mostly alternative choices play to that of binary choices play.

Further observations on individual runs for the experiment with reputation reveal strategies evolving to play binary choices, i.e., full cooperation and full defection. We plot the sampling distribution of choices played by the population to show this in Fig. 13. For example, in all runs that evolved to cooperation, we can observe the population evolving to play mostly binary choices in less than 100 generations (Fig. 13). The higher frequency of full cooperation play compared with that of full defection play indicates a cooperative outcome for the population. For runs that evolved to defection, the plots show the opposite.

Although plots such as in Fig. 13 can reveal how frequent choices are played in a generation and how the distribution changes throughout the evolutionary process, they do not provide information on strategy behaviors and how they change during evolution. As in the previous section, we observe the

Fig. 14. Properties of most fit evolved strategies in the run that evolved to cooperation for reputation estimation. The top graph plots the choices played for the first move across generations. The middle graph plots the choices played if previous move is full cooperation across generations. The lower graph plots the choices played if previous move is full defection across generations.

Fig. 15. Properties of most fit evolved strategies in the run that evolved to cooperation for opponent with good reputation. The top graph plots the choices played for the first move across generations. The middle graph plots the choices played if previous move is full cooperation across generations. The lower graph plots the choices played if previous move is full defection across generations.

evolution of three properties of the evolved and most fit strategies (the first move, the response to mutual cooperation, and the response to mutual defection) to facilitate our understanding of how strategy behaviors change during evolution that lead to binary choices of play.

Our observation on runs that resulted in cooperation reveals that strategies evolved to be naive, highly cooperative players for the reputation estimation stage (Fig. 14), e.g., nice (top graph), respond with full cooperation regardless of whether previous moves are mutual cooperation (middle graph), or mutual defection (bottom graph). For the second stage that considers reputation, we observe that the same strategies evolved to play highly cooperative choices (Fig. 15) if the opponent has a good reputation (equal and higher reputation scores). However, for the case of opponents with bad reputations (lower reputation scores), the same strategies (Fig. 16) evolved to be unforgiving, defective players, starting with full defection play

Fig. 16. Properties of most fit evolved strategies in the run that evolved to cooperation for opponent with bad reputation. The top graph plots the choices played for the first move across generations. The middle graph plots the choices played if previous move is full cooperation across generations. The lower graph plots the choices played if previous move is full defection across generations.

for the first move (top graph), and then continue to respond with full defection regardless of the case of mutual cooperation (middle graph) or mutual defection (bottom graph) in the previous moves.

3) *How Reputation Helps: Strategy Responding Mainly to Reputation Scores:* Although Figs. 14–16 indicate the evolution of strategies responding mostly with binary choices, they do not provide information as to how evolution of cooperation is achieved where strategies are able to engage in full cooperation play with each other. We inspect the actual behavioral responses of strategies to determine how cooperative outcomes are achieved.

We start the analysis with behavioral responses of the final generation's most fit evolved strategy. Observations during the reputation estimation stage show that the strategy plays nice and continues to be highly cooperative (Fig. 17) in order to obtain a good reputation since the score is calculated from the distribution of choices played. During the second stage that considers reputation scores, the strategy plays “all cooperate” with the opponent that has a good reputation (Fig. 18), but plays “all defect” against opponent with bad reputation (Fig. 19). In effect, the strategy appears to adopt a “discriminatory” play based mostly on reputation scores input whereby it cooperates if the opponent has a good reputation, but defect if the opponent has a bad reputation. These results suggest that strategies evolved to use only reputation for behavioral response, and that inputs for previous moves are effectively superfluous.

In addition, we also observe that not only the most fit evolved strategy of the population adopts a “discriminatory” play, but that this strategy dominates the population. A “discriminatory” strategy is successful because it is able to engage in full cooperation play throughout the entire game with similar strategies, thus obtaining high payoffs. Here, reputation acts as a sort of signaling mechanism for strategies to engage in mutual cooperation. High reputation scores reflect a strategy's willingness to engage in highly cooperative play. Lower reputation scores are indicative of a strategy's unwillingness to engage in highly

Fig. 17. Most fit evolved strategy's behavioral response during reputation estimation stage at the end of an evolutionary run that resulted in highly cooperative play. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 63. Note that the triangular ridge of full defection responses is a result of evolution of earlier strategies that played intermediate choices. At this point, the strategy has effectively evolve to play "all cooperate" since it starts with full cooperation, and continues to cooperate regardless of the opponent's response. The triangular regions are never accessed during a behavioral exchange.

Fig. 18. Most fit evolved strategy's behavioral response during the IPD with reputation stage at the end of an evolutionary run that resulted in highly cooperative play. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 63. Note for the case that the opponent has a good reputation, the figure shows that the strategy effectively plays "all cooperate."

Fig. 19. Most fit evolved strategy's behavioral response during the IPD with reputation stage at the end of an evolutionary run that resulted in highly cooperative play. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 63. Note for the case that the opponent has a bad reputation, the figure shows that the strategy effectively plays "all defect."

cooperative play. As such, by comparing reputation scores, a "discriminatory" strategy knows beforehand which strategy to

trust by playing "all cooperate," and which strategy it should not trust by uncompromisingly playing "all defect."

These results suggest that strategies have evolved to consider reputation scores only despite having inputs for both reputation scores and previous moves. Our examination of the coevolutionary learning model reveals that there is higher selective pressure during evolution on traits connected to reputation scores when a small p_r value that is less than 0.5 is used, and that the "discriminatory" play can be traced to the substantially higher connection weights of the neural networks for the fifth input (reputation) compared with the other four inputs for previous moves. This higher selective pressure on reputation inputs can be explained by considering that there are more games where a strategy use reputation (i.e., second stage) that contribute to its fitness compared to games without considering reputation (i.e., first stage) when p_r is less than 0.5. As such, there is a feedback to evolution to place higher selective pressure on traits associated with reputation.

We conducted an experiment that used $p_r = 0.5$ to further clarify this. We note that there is no distinction between games with and without reputation in their contribution to a strategy's fitness because both games are played in similar numbers on average. We observe in runs with cooperative outcomes that final evolved strategies are not fully "discriminatory" with respect to the opponent's reputation, as shown by Figs. 18 and 19.

At this point, we note that our results of strategies evolving to consider only reputation corresponds to the results, and also justify the motivation of previous studies [21], [27] that modeled behavioral responses based solely on reputation for input and decision-making. As mentioned by the same authors in [20], the mechanism of direct reciprocity was removed from their experiments both for simplicity and also to focus on how the mechanism of indirect reciprocity can result with cooperation. However, they also pointed out that indirect reciprocity is a result of direct reciprocity that occurs between other individuals. Here, we made this process that connects these two mechanisms explicit: reputation scores of strategies are estimated from choices that they made in a small number of interactions that do not involve reputation in the population, which are then used in interactions that involve reputation with other remaining strategies. From these two mechanisms, the coevolutionary process produces strategies with cooperative behaviors.

4) *Does Maintaining a Good Reputation Pay?*: A question that immediately follows is whether it pays for strategies to maintain good reputation. This is important because adding another input to the decision-making process, i.e., reputation, is more than just introducing an extra dimension to the behavioral response space. Instead, due to reputation scores being calculated from choices that strategies make during their interactions, the additional input for reputation introduces a subtle and complex relationship between responding to the current interaction (gaining fitness from the current game) and responding for future interactions (gaining probable fitness from other members of the population by maintaining favorable reputation scores).

Recall that strategies consider both their own reputation scores and that of the opponent's. A strategy considers its opponent having a good reputation if the opponent's reputation score is equal to or higher than its reputation score. We observed

Fig. 20. Evolutionary dynamics for a sample run that evolved to cooperative play. The top graph plots the average population payoff across generations. The lower graph plots the reputation score of the most fit evolved strategy of the population across generations.

that in all our runs that evolved cooperative play, the most fit strategy of the population has a high reputation score. We plot Fig. 20 for a typical run that evolved cooperative play, which allows the comparison on how the reputation score of successful strategies (i.e., the evolved and most fit strategy in each generation) changes during evolution (bottom graph) with how the state of cooperation (using the average population payoff) in the population changes (top graph). The figure shows that the rise of the average population payoff coincides with the rise of the reputation score of the most fit evolved strategy. For runs that ended with defection play, we observe that the reputation of most fit evolved strategy fluctuates between -1 (obtained by playing full defection) and 0 (obtained by not being sampled for the reputation estimation stage).

From our experiment results, we observed that not only the evolved and most fit strategy maintains a high reputation score of one, which is the highest possible value, but that the population consists of strategies with high reputation scores of one. This is achieved when these strategies play more full cooperation levels (e.g., broader surfaces for full cooperation in Fig. 17) during the reputation estimation stage. This is because the reputation score of a strategy is calculated based on choices that it played. Strategies that are more likely to respond with full cooperation regardless of the pair of moves made in the previous round (broader surfaces for full cooperation) are more likely to obtain higher reputation scores.

With a high reputation score, the strategy is able to elicit mutual cooperative play from similar strategies with comparable reputation scores, but defect against other strategies with lower reputation scores because they are more likely not to engage in highly cooperative play. In this way, reputation helps with cooperation because it discourages strategies from playing other alternative choices that lead to lower cooperation levels, especially in the situation of games with more choices and shorter durations.

Unlike previous studies that focused only on the mechanism of indirect reciprocity [21], [27], one concern with the approach

here is the possibility of having cheaters, i.e., strategies that obtain high reputation score to be perceived as having good reputation so that they can exploit “naïve” cooperators. However, our implementation discourages cheaters. That is, although functionally neural networks can be rewired to adopt a “cheating” play, they are not evolved (results from experiment show that for runs with cooperative outcomes, strategies obtain high reputation scores by being more cooperative). One reason for this observation is that a “cheater” does not do well against other “cheaters” (i.e., self-play). Furthermore, “cheaters” themselves are exploitable by other “cheaters” that play a lower cooperation level where further evolution will drive the population to mutual defection.

V. THE IMPACT OF DIFFERENT REPUTATION ESTIMATIONS

A. Limitations of the Two-Stage Procedure for Reputation Estimation

We have shown earlier that reputation can help the evolution of cooperation in the IPD with more choices and also for the case of short game durations. However, the two-stage procedure might not be a realistic implementation for two reasons.

First, the two-stage procedure limits the setting of the number of games a strategy can play with others in the population for reputation estimation. Earlier, our results show that starting from a small p_r value, increasingly higher values of p_r have a positive and significant impact on the evolution of cooperation up to a certain point (e.g., $p_r = 0.15$ in Tables IX and X). With higher p_r values, each strategy is tested against more opponents, and as such, the reputation score better reflects its behavior that is crucial for games in the second stage. However, further increase to p_r does not lead to more cooperative outcomes because there are less games in the second stage to be played for which strategies can take advantage of to engage in highly cooperative plays.

Second, the reputation scores of strategies are estimated anew every generation. There is no memory of games from previous generations in the calculation of reputation scores. As such, the reputation score that is obtained can be misleading because the procedure in updating the reputation score is not cumulative. For example, a strategy that obtains a high reputation score because of interactions with cooperative opponents can obtain a low reputation score in the following generation because it defects against random opponents that play low cooperation levels only.

For the second limitation, incorporating the memory of games from previous generations in calculating reputation scores may not be trivially addressed. We consider the approach that updates reputation scores using raw scores from previous generations. For example, we can obtain the new reputation score R_{new} , by adding weighted reputation score of the current generation R_{current} , and that of the previous generation's, R_{previous}

$$\begin{aligned} R_{\text{new}} &= \beta * R_{\text{previous}} + \alpha * R_{\text{current}} \\ &= R_{\text{previous}} + \alpha * (R_{\text{current}} - R_{\text{previous}}) \end{aligned} \quad (4)$$

where $\alpha + \beta = 1$, and R_{new} is always in the range of $[-1, 1]$. Since R_{previous} is unknown at the start of the evolution, R_{new} is similar to the original calculation given by (2). Equation (4)

TABLE XI

RESULTS OF EXPERIMENTS THAT INCLUDE REPUTATION AND MEMORY OF REPUTATION SCORES OF THE PREVIOUS GENERATION. *RepW1* REFERS TO EXPERIMENTS THAT USED EQUATION (4) FOR UPDATING REPUTATION SCORES. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “MEAN” INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. “STD DEV” INDICATES THE STANDARD DEVIATION. “MAX” INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. “MIN” INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE *p*-VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE EXPERIMENT WITHOUT REPUTATION AND THAT OF EXPERIMENTS WITH REPUTATION AND OTHER α VALUES

Experiment	Mean $\pm$ Std Dev	Max	Min	<i>p</i> -value
<i>NoRep</i>	1.37 $\pm$ 0.84	3.87	1.00	
<i>RepW1</i> , $\alpha = 0.9$	2.19 $\pm$ 1.30	3.94	1.00	0.0998
<i>RepW1</i> , $\alpha = 0.7$	2.34 $\pm$ 1.27	4.00	1.00	0.0176 [†]
<i>RepW1</i> , $\alpha = 0.5$	2.25 $\pm$ 1.28	4.00	1.00	0.0248 [†]

[†] The *p*-value indicates a statistically significant result for a 0.05 level of significance that is considered.

TABLE XII

RESULTS OF EXPERIMENTS THAT INCLUDE REPUTATION AND MEMORY OF REPUTATION SCORES OF THE PREVIOUS GENERATION. *RepW1* REFERS TO EXPERIMENTS THAT USED EQUATION (4) FOR UPDATING REPUTATION SCORES. RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “ $No \leq 1.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFECTION PLAY. “ $1.5 < No < 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. “ $No \geq 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Experiment	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
<i>NoRep</i>	25	4	1
<i>RepW1</i> , $\alpha = 0.9$	16	6	8
<i>RepW1</i> , $\alpha = 0.7$	14	8	8
<i>RepW1</i> , $\alpha = 0.5$	15	8	7

is applied from the second generation onwards. We assume that the calculation for the current generation is more important than the previous generation’s, e.g., $\alpha \geq 0.5$.

Tables XI and XII summarize experimental results. Results indicate that the impact of incorporating the memory of games from previous generations using raw reputation scores is unclear since there is no improvement in the number of runs evolved to cooperation. With the addition of memory, we seek to increase the accuracy of the procedure estimating the behavior of a strategy based on games played with other individuals by accumulating the scores from previous generations. When (4) is applied throughout the evolutionary process, the estimations that are several generations back are effectively represented as a single value, e.g., R_{previous} . Although the equations model the effect of reputation scores that are further back in the evolutionary process having lesser importance (e.g., $\beta \leq \alpha$), the procedure that is based on using raw reputation scores instead of actual choices played might be inaccurate.

B. More Accurate Reputation Estimation

Due to the limitations of the two-stage procedure for calculating reputation scores, we propose an alternative implementation for the IPD game that considers previous moves and reputation simultaneously. We address the first limitation by considering an approach with a single-stage procedure, whereby in

every generation, all strategies in the population compete in the *n*-choice IPD with reputation. This also removes adding an additional parameter to the coevolutionary learning model (e.g., p_r in the original, two-stage procedure).

The second limitation that involves incorporating memory of games from previous generations can be addressed by having each strategy storing its frequency distribution of choices played, $f(x)$, and updating it for every game played. Surviving strategies also carry forward their distributions to the next generation. Although this increases the memory requirement, reputation scores can be accurately determined from choices played by applying (2).

More importantly, with the new procedure, we can study how different interpretations on estimating a strategy’s reputation can affect on the evolutionary outcome. We focus on the issue of the accuracy of reputation estimation, i.e., how well a strategy’s reputation score reflects its behavior based on games played with other individuals from the procedure. In particular, we investigate how accuracy is related to the procedure of incorporating memory of games from previous generations to calculate reputation scores and how frequently reputation scores are updated. Although the procedures may appear contrary to the original motivation that reputation reflects the behavior of a strategy based on a small number of games played in the past, we note that what is more important is that the update to the reputation score will still affect a strategy’s interactions with the majority of future opponents.

We note here that the procedure for selection and variation of strategies remain the same. As for behavioral responses of the representation, strategies now consider, in addition to previous moves, the case of whether their opponents have good or bad reputations. That is, the fifth input of the neural network now takes either +1 or −1. It does not take in 0 anymore, which was originally used in the reputation estimation stage of the two-stage procedure.

1) *Update Reputation After Every Generation:* The coevolutionary learning model for the alternative reputation estimation differs from the model described in Section IV in how a reputation score is determined. In particular, each strategy now also stores its frequency distribution of choices played, $f(x)$, $x = \{x_1, x_2, \dots, x_n\}$, which is initialized to 0 for all x . The offspring inherits the parent’s frequency distribution of choices played, $f(x)$. Reputation score for each strategy is also calculated using (2) at the start of every generation, which is unchanged once it is calculated. At the start of the evolutionary process, due to the initialization of strategies’ $f(x)$ s, all strategies are arbitrarily assigned random reputation scores from a uniform distribution in the range of $[-1, 1]$. Note that given how reputation is assigned based on comparison of reputation scores (a strategy has good reputation if it has the same reputation score as its partner’s), a uniform assignment such as initializing all strategies with zero reputation score is not implemented because all strategies will only access their behavioral maps corresponding to good reputation opponent only.

Tables XIII and XIV summarize the experimental results. Comparing the population average payoffs of individual runs of the experiment without reputation (*NoRep*) and the experiment with (*RepA1*), results show a statistically significant difference

TABLE XIII

RESULTS OF EXPERIMENTS THAT USE ALTERNATIVE REPUTATION ESTIMATION BASED ON FREQUENCY DISTRIBUTION OF CHOICES PLAYED, $f(x)$, THAT IS ACCUMULATED FOR SURVIVING STRATEGIES THROUGHOUT EVOLUTION. *RepA1* EVOLVED STRATEGIES TO PLAY THE GAME OF IPD WITH MORE CHOICES AND REPUTATION DIRECTLY. EXPERIMENTS ARE ALSO CONDUCTED TO DETERMINE THE EVOLUTIONARY OUTCOME IN SITUATIONS WHEREBY SOME GAMES ARE NOT REGISTERED FOR UPDATING $f(x)$ BY SETTING A PROBABILITY, p_f . RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “MEAN” INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. “STD DEV” INDICATES THE STANDARD DEVIATION. “MAX” INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. “MIN” INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE p -VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE *NoRep* EXPERIMENT WITH THAT OF OTHER *RepA1* EXPERIMENTS. STATISTICAL TESTS BETWEEN *Rep* AND *RepA1* (NOT SHOWN IN TABLE) SHOWED NO STATISTICAL SIGNIFICANT DIFFERENCE FOR A 0.05 LEVEL OF SIGNIFICANCE THAT IS CONSIDERED

Experiment	Mean $\pm$ Std Dev	Max	Min	p -value
<i>NoRep</i>	1.37 $\pm$ 0.84	3.87	1.00	
<i>RepA1</i>	2.63 $\pm$ 1.42	4.00	1.00	0.0007 [†]
<i>RepA1</i> , $p_f = 0.1$	2.17 $\pm$ 1.38	4.00	1.00	0.0281 [†]
<i>RepA1</i> , $p_f = 0.2$	2.33 $\pm$ 1.45	4.00	1.00	0.0396 [†]
<i>RepA1</i> , $p_f = 0.3$	2.32 $\pm$ 1.39	4.00	1.00	0.0127 [†]
<i>RepA1</i> , $p_f = 0.4$	2.40 $\pm$ 1.43	4.00	1.00	0.0374 [†]
<i>RepA1</i> , $p_f = 0.5$	2.40 $\pm$ 1.41	4.00	1.00	0.0218 [†]
<i>RepA1</i> , $p_f = 0.6$	2.55 $\pm$ 1.44	4.00	1.00	0.0037 [†]
<i>RepA1</i> , $p_f = 0.7$	2.24 $\pm$ 1.38	4.00	1.00	0.0647
<i>RepA1</i> , $p_f = 0.8$	2.36 $\pm$ 1.47	4.00	1.00	0.0400 [†]
<i>RepA1</i> , $p_f = 0.9$	2.25 $\pm$ 1.43	4.00	1.00	0.0294 [†]

[†] The p -value indicates a statistically significant result for a 0.05 level of significance that is considered.

(Table XIII). Although there is no statistically significant difference between the alternative reputation estimation approach proposed here (*RepA1*) and that of the original two-stage procedure (*Rep*), the alternative approach did result with more cooperative outcomes (e.g., 15 for *RepA1* compared with eight for *Rep*, as shown in Table XIV).

In addition, we conducted experiments to determine the evolutionary outcome in situations where not all games are used to update a strategy’s frequency distribution of choices played, $f(x)$, to model the situation where choices that a strategy played in some games are forgotten or unknown to other strategies for reputation estimation. We used a simple implementation where every game has a probability of not being registered for updating $f(x)$, p_f , and that all the choices played in that particular game are forgotten. Results in Tables XIII and XIV show that evolution of cooperation is still possible. Although the number of cooperative outcomes is reduced given the decrease in the accuracy of the estimation when not all games are registered in updating $f(x)$, there are still more cooperative outcomes in comparison with the original two-stage procedure (*Rep*). There are no statistically significant differences in results when comparing *RepA1* experiment with other experiments with various p_f (results of statistical tests are omitted in Table XIII).

In terms of how strategy behaviors evolve, our analysis reveals strategies evolving to play mostly binary choices of either

TABLE XIV

RESULTS OF EXPERIMENTS THAT USE ALTERNATIVE REPUTATION ESTIMATION BASED ON FREQUENCY DISTRIBUTION OF CHOICES PLAYED, $f(x)$, THAT IS ACCUMULATED FOR SURVIVING STRATEGIES THROUGHOUT EVOLUTION. *RepA1* EVOLVED STRATEGIES TO PLAY THE GAME OF IPD WITH MORE CHOICES AND REPUTATION DIRECTLY. EXPERIMENTS ARE ALSO CONDUCTED TO DETERMINE THE EVOLUTIONARY OUTCOME IN SITUATIONS WHEREBY SOME GAMES ARE NOT REGISTERED FOR UPDATING $f(x)$ BY SETTING A PROBABILITY, p_f . RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. “ $No \leq 1.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFLECTION PLAY. “ $1.5 < No < 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. “ $No \geq 3.5$ ” INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Experiment	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
<i>NoRep</i>	25	4	1
<i>RepA1</i>	12	3	15
<i>RepA1</i> , $p_f = 0.1$	17	2	11
<i>RepA1</i> , $p_f = 0.2$	16	2	12
<i>RepA1</i> , $p_f = 0.3$	15	4	11
<i>RepA1</i> , $p_f = 0.4$	15	2	13
<i>RepA1</i> , $p_f = 0.5$	15	2	13
<i>RepA1</i> , $p_f = 0.6$	14	2	15
<i>RepA1</i> , $p_f = 0.7$	16	1	10
<i>RepA1</i> , $p_f = 0.8$	16	0	14
<i>RepA1</i> , $p_f = 0.9$	17	0	13

Fig. 21. Most fit evolved strategy’s behavioral response if the opponent has a good reputation at the end of an evolutionary run that resulted in highly cooperative play for the experiment that used the alternative reputation estimation based on choices played in previous generations *RepA1*. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 63.

full cooperation or full defection (Figs. 21 and 22). Closer observations of the strategies’ behavioral responses reveal highly cooperative strategies, both for opponents with good and bad reputation. However, with the exception of good reputation opponents (Fig. 21), strategies will respond with very low cooperation levels against bad reputation opponents that played very low cooperation levels (Fig. 22).

Unlike evolved strategies from the two-stage procedure, strategies here did not evolve to purely “discriminatory” play based on reputation alone. One likely explanation as to why “discriminatory” strategies are evolved using the two-stage procedure, but not with the single-stage procedure, is to note that in the two-stage procedure, the fitness of a strategy is

Fig. 22. Most fit evolved strategy's behavioral response if the opponent has a bad reputation at the end of an evolutionary run that resulted in highly cooperative play for the experiment that used the alternative reputation estimation based on choices played in previous generations *RepA1*. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 43.

TABLE XV

RESULTS OF EXPERIMENTS THAT USE A SECOND ALTERNATIVE REPUTATION ESTIMATION BASED ON FREQUENCY DISTRIBUTION OF CHOICES PLAYED, $f(x)$, THAT IS ACCUMULATED FOR SURVIVING STRATEGIES THROUGHOUT EVOLUTION. WITH *RepA2*, REPUTATION SCORE FOR A STRATEGY IS UPDATED EVERY TIME $f(x)$ IS UPDATED FOR EVERY GAME PLAYED IN A GENERATION. EXPERIMENTS ARE ALSO CONDUCTED TO DETERMINE THE EVOLUTIONARY OUTCOME IN SITUATIONS WHEREBY SOME MOVES IN A GAME ARE NOT REGISTERED FOR UPDATING $f(x)$ BY SETTING A PROBABILITY p_f . RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. "MEAN" INDICATES THE AVERAGE OVER 30 RUNS OF POPULATION AVERAGE PAYOFF. "STD DEV" INDICATES THE STANDARD DEVIATION. "MAX" INDICATES THE HIGHEST POPULATION AVERAGE PAYOFF. "MIN" INDICATES THE LOWEST POPULATION AVERAGE PAYOFF. THE p -VALUE IS OBTAINED FROM A PAIRED WILCOXON SIGNED RANK TEST THAT COMPARES THE POPULATION AVERAGE PAYOFFS OF INDIVIDUAL RUNS FOR THE *NoRep* EXPERIMENT WITH THAT OF OTHER *RepA2* EXPERIMENTS. STATISTICAL TESTS BETWEEN *Rep* AND *RepA2* (NOT SHOWN IN TABLE) SHOWED STATISTICAL SIGNIFICANT DIFFERENCE FOR A 0.05 LEVEL OF SIGNIFICANCE THAT IS CONSIDERED

Experiment	Mean $\pm$ Std Dev	Max	Min	p -value
<i>NoRep</i>	1.37 $\pm$ 0.84	3.87	1.00	
<i>RepA2</i>	3.24 $\pm$ 1.06	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.1$	3.14 $\pm$ 1.13	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.2$	3.61 $\pm$ 0.84	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.3$	3.27 $\pm$ 1.03	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.4$	3.34 $\pm$ 1.06	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.5$	2.94 $\pm$ 1.08	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.6$	2.95 $\pm$ 1.09	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.7$	2.95 $\pm$ 1.24	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.8$	3.20 $\pm$ 1.14	4.00	1.00	< 0.0001 [†]
<i>RepA2</i> , $p_f = 0.9$	3.34 $\pm$ 1.00	4.00	1.00	< 0.0001 [†]

[†] The p -value indicates a statistically significant result for a 0.05 level of significance that is considered.

the average payoff from games with (first stage) and without (second stage) reputation. Given that the number of games from the two stages are not equal, i.e., more games for the second stage with reputation, one would expect a selective pressure on the input for reputation. However, for the single-stage approach, there is no such selective pressure for the input for reputation because all games are played with reputation. We did not observe markedly higher connection weights for the input

TABLE XVI

RESULTS OF EXPERIMENTS THAT USE A SECOND ALTERNATIVE REPUTATION ESTIMATION BASED ON FREQUENCY DISTRIBUTION OF CHOICES PLAYED, $f(x)$, THAT IS ACCUMULATED FOR SURVIVING STRATEGIES THROUGHOUT EVOLUTION. WITH *RepA2*, REPUTATION SCORE FOR A STRATEGY IS UPDATED EVERY TIME $f(x)$ IS UPDATED FOR EVERY GAME PLAYED IN A GENERATION. EXPERIMENTS ARE ALSO CONDUCTED TO DETERMINE THE EVOLUTIONARY OUTCOME IN SITUATIONS WHEREBY SOME MOVES IN A GAME ARE NOT REGISTERED FOR UPDATING $f(x)$ BY SETTING A PROBABILITY p_f . RESULTS TAKEN AT THE END OF THE COEVOLUTIONARY LEARNING RUNS. " $No \leq 1.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL DEFECTION PLAY. " $1.5 < No < 3.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO ALTERNATE IN THEIR PLAY OR TO PLAY INTERMEDIATE CHOICES. " $No \geq 3.5$ " INDICATES THE NUMBER OF RUNS WHERE THE POPULATIONS EVOLVED TO MUTUAL COOPERATION PLAY

Experiment	$No \leq 1.5$	$1.5 < No < 3.5$	$No \geq 3.5$
<i>NoRep</i>	25	4	1
<i>RepA2</i>	5	5	20
<i>RepA2</i> , $p_f = 0.1$	5	7	18
<i>RepA2</i> , $p_f = 0.2$	2	1	27
<i>RepA2</i> , $p_f = 0.3$	4	6	20
<i>RepA2</i> , $p_f = 0.4$	3	5	22
<i>RepA2</i> , $p_f = 0.5$	6	9	15
<i>RepA2</i> , $p_f = 0.6$	6	9	15
<i>RepA2</i> , $p_f = 0.7$	7	5	18
<i>RepA2</i> , $p_f = 0.8$	6	4	20
<i>RepA2</i> , $p_f = 0.9$	4	5	21

corresponding to reputation (the same observation is also made for the two-stage procedure when strategies played a lower number of games for the second stage that involve reputation).

2) *Update Reputation After Every Game*: Experiments conducted here build on top of those conducted earlier in that not only memory of games from previous generations are incorporated to calculate reputation scores, but that reputation scores are updated after every game played. Results are summarized in Tables XV and XVI. Results in Tables XVI and XIV suggest that updating reputation scores on the fly as each game is played has a positive impact on cooperative outcomes as compared with updating scores only after all games have been played. There is a statistically significant difference in the results between the alternative reputation approach used here that updates reputation scores after every game (*RepA2*) and the original two-stage procedure (*Rep*) (results of statistical tests are omitted in Table XV). Our analysis indicates evolved strategies with similar mutually cooperative responses compared with that shown earlier by Figs. 21 and 22. In addition, we conducted experiments to determine whether the evolutionary outcome is affected if not all moves in a game are used to update a strategy's frequency distribution of choices, $f(x)$. Results in Table XVI show that the setting of p_f (probability that a move in a game is forgotten) does not seem to affect greatly the number of cooperative outcomes. There are no statistically significant differences when comparing *RepA2* experiment with the other experiments with various p_f (results of statistical tests are omitted in Table XV).

As for the evolution of strategy behaviors, we observed similar evolved behaviors when comparing results of the *RepA2* experiment with that of *RepA1*. Strategies evolve to

Fig. 23. Most fit evolved strategy's behavioral response if the opponent has a good reputation at the end of an evolutionary run that resulted in highly cooperative play for the experiment that used the alternative reputation estimation based on choices played in previous generations *Rep.A1*. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 63.

Fig. 24. Most fit evolved strategy's behavioral response if the opponent has a bad reputation at the end of an evolutionary run that resulted in highly cooperative play for the experiment that used the alternative reputation estimation based on choices played in previous generations *Rep.A1*. 0 indicates full defection, while 63 indicates full cooperation. This strategy started with a move of 27.

play mostly binary choices of either full cooperation or full defection (Figs. 23 and 24). They are highly cooperative with both good and bad reputation opponents that are cooperative. Although strategies are not purely “discriminatory” based on reputation alone, they are less tolerant (e.g., response with full defection for opponents that played lower cooperation levels) to bad reputation opponents (Fig. 22), as compared with good reputation opponents (Fig. 21). That is, strategies do not discriminate based entirely on reputation alone, as further shown in Figs. 25 and 26 that show how strategy behavior changes during evolution (we also did not note any significant difference between input connection weights of the neural networks associated with previous moves and reputation). Instead, discrimination is more subtle and more dependent on how interactions unfold (previous moves and reputation are considered simultaneously).

Results here suggest the importance of accumulating the memory of games from the previous generation to calculate reputation scores and having them updated after every game, which leads to further increase in the accuracy of reputation estimation, for further evolution of cooperation. This is different compared with the two-stage procedure where the incorporation

Fig. 25. Properties of most fit evolved strategies in the run that evolved to cooperation for opponent with good reputation for the experiment that used alternative reputation estimation based on choices played in previous generations *Rep.A2*. The top graph plots the choices played for the first move across generations. The middle graph plots the choices played if previous move is full cooperation across generations. The lower graph plots the choices played if previous move is full defection across generations.

Fig. 26. Properties of most fit evolved strategies in the run that evolved to cooperation for opponent with bad reputation for the experiment that used alternative reputation estimation based on choices played in previous generations *Rep.A2*. The top graph plots the choices played for the first move across generations. The middle graph plots the choices played if previous move is full cooperation across generations. The lower graph plots the choices played if previous move is full defection across generations.

of memory based on raw reputation scores has no significant and positive impact on the evolutionary outcome. However, we note that for comparisons between the two single-stage alternatives and the original two-stage procedure, only a significant and positive impact on evolutionary outcomes (e.g., cooperation) is obtained when reputation is updated after every game, not when it is updated after every generation. Given that every decision-making for games involves previous moves and reputation scores simultaneously for the single-stage procedure, further increase in the accuracy of reputation estimation to affect the evolutionary outcome will depend on how frequently reputation is updated.

VI. CONCLUSION

This paper presents a series of investigations aimed to further understand why coevolutionary learning is less likely to produce cooperative outcomes as the number of available choices to play in the IPD game increases. This paper focuses on finding out why more choices lead to defection outcomes in the coevolutionary learning of IPD strategies, and how subsequent analysis of the results leads to devising a solution for the coevolutionary learning of cooperative behaviors.

Based on our studies, it appears that when there are more choices, strategies have more opportunities to exploit others by playing lower cooperation levels. We note that with more choices, strategies are less able to resolve the intention of an intermediate choice, i.e., whether it is a signal to engender further cooperation or a subtle exploitation. Strategies adapt to play defection because they provide higher payoffs in the short-term when they are unable to resolve the intention of opponents. In effect, strategies are less likely to learn highly cooperative plays as cooperative plays are sampled less often compared with plays with lower cooperation levels.

Based on this result, and motivated from the view that complex human interactions are not limited to only direct interactions, we investigate other mechanisms known to promote cooperation without depending on repeated encounters (direct reciprocity). One such mechanism is indirect reciprocity. The mechanism of indirect reciprocity can be explained in terms of reputation, whereby cooperative behaviors between two individuals depend on their prior behaviors with other individuals. Reputation is incorporated into the IPD with more choices by adding an additional input for reputation to existing inputs for previous moves in a strategy's behavioral response. A strategy's reputation is determined relative to the current interacting pair's reputation scores, which in turn are determined based on choices played by them to other strategies prior to the current game. The resulting IPD game with more choices and reputation that we consider here is very different from other studies. The coevolutionary learning of cooperative behaviors are possible and more likely when strategies use reputation as a mechanism to estimate behaviors of future partners prior to interactions. Through reputation, cooperative strategies are able to elicit mutual cooperation play right from the start of the interaction, which is essential for mutual cooperation in the IPD game with more choices and short game durations.

A strategy's reputation can be estimated in many different ways. We investigated the impact of different implementations of reputation estimation on the evolution of cooperation, focusing on one major issue: the accuracy of reputation estimation. Experiments show that incorporating memory of games from previous generations using actual choices played from games, not raw reputation scores, leads to a higher accuracy in reputation estimation that has a positive impact on the evolution of cooperation. Further evolution of cooperation is obtained when one considers an almost instantaneous feedback of a strategy's behavior to others in the population by updating the reputation score for every game played.

The framework introduced here, which considers both direct and indirect interactions, provides avenues of study for many

issues on the coevolutionary learning of behaviors. One issue that requires further study is noise and how it can affect the evolutionary outcome. This is because noise can affect the process of estimating strategy behaviors. In addition, the issue of the robustness of evolved strategies that considers both previous moves and reputation should be investigated. Further study is also required to understand the impact of strategies that cheat, i.e., obtained good reputation, but later evolved to exploit "naive" (in terms of observing reputation) strategies.

ACKNOWLEDGMENT

The authors would like to thank Dr. D. B. Fogel for his role as the acting Editor-in-Chief for this paper. They would also like to thank the three anonymous reviewers for their constructive comments, which have helped to improve this paper significantly. They are grateful to The Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), School of Computer Science, The University of Birmingham, for providing computing support in running the experiments.

REFERENCES

- [1] R. Axelrod, "Effective choice in the prisoner's dilemma," *J. Conflict Resolution*, vol. 24, no. 1, pp. 3–25, Mar. 1980.
- [2] —, "More effective choice in the prisoner's dilemma," *J. Conflict Resolution*, vol. 24, no. 3, pp. 379–403, Sep. 1980.
- [3] —, *The Evolution of Cooperation*. New York: Basic Books, 1984.
- [4] —, "The evolution of strategies in the iterated prisoner's dilemma," in *Genetic Algorithms and Simulated Annealing*, L. D. Davis, Ed.. New York: Morgan Kaufmann, 1987, ch. 3, pp. 32–41.
- [5] D. B. Fogel, "The evolution of intelligent decision making in gaming," *Int. J. Cybern. Syst.*, vol. 22, pp. 223–236, 1991.
- [6] —, "Evolving behaviors in the iterated prisoner's dilemma," *Evol. Comput.*, vol. 1, no. 1, pp. 77–97, 1993.
- [7] —, "On the relationship between the duration of an encounter and the evolution of cooperation in the iterated prisoner's dilemma," *Evol. Comput.*, vol. 3, no. 3, pp. 349–363, 1996.
- [8] B. A. Julstrom, "Effects of contest length and noise on reciprocal altruism, cooperation, and payoffs in the iterated prisoner's dilemma," in *Proc. 7th Int. Conf. Genetic Algorithms*, 1997, pp. 386–392.
- [9] P. Darwen and X. Yao, "On evolving robust strategies for iterated prisoner's dilemma," *Progress Evol. Comput.*, ser. Lecture Notes in Artificial Intelligence, vol. 956, pp. 276–292, 1995.
- [10] J. Batali and P. Kitcher, "Evolution of altruism in optional and compulsory games," *J. Theor. Biol.*, vol. 175, pp. 161–171, 1995.
- [11] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Science*, vol. 211, pp. 1390–1396, 1981.
- [12] K. Lindgren, "Evolutionary phenomena in simple dynamics," in *Artif. Life II*, C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, Eds. Reading, MA: Addison-Wesley, 1991, pp. 295–312.
- [13] P. G. Harrald and D. B. Fogel, "Evolving continuous behaviors in the iterated prisoner's dilemma," *BioSystems: Special Issue on the Prisoner's Dilemma*, vol. 37, pp. 135–145, 1996.
- [14] P. Darwen and X. Yao, "Does extra genetic diversity maintain escalation in a co-evolutionary arms race," *Int. J. Knowledge-Based Intell. Eng. Syst.*, vol. 4, no. 3, pp. 191–200, Jul. 2000.
- [15] —, "Why more choices cause less cooperation in iterated prisoner's dilemma," in *Proc. 2001 Congr. Evol. Comput.*, 2001, pp. 987–994.
- [16] —, "Co-evolution in iterated prisoner's dilemma with intermediate levels of cooperation: Application to missile defense," *Int. J. Comput. Intell. Appl.*, vol. 2, no. 1, pp. 83–107, 2002.
- [17] L. M. Wahl and M. A. Nowak, "The continuous prisoner's dilemma: I. linear reactive strategies," *J. Theor. Biol.*, vol. 200, no. 3, pp. 307–321, 1999.
- [18] M. R. Frean and E. R. Abraham, "A voter model of the spatial prisoner's dilemma," *IEEE Trans. Evol. Comput.*, vol. 5, no. 2, pp. 117–121, Apr. 2001.
- [19] S. Y. Chong and X. Yao, "Behavioral diversity, choices, and noise in the iterated prisoner's dilemma," *IEEE Trans. Evol. Comput.*, vol. 9, no. 6, pp. 540–551, Dec. 2005.

- [20] M. A. Nowak and K. Sigmund, "The dynamics of indirect reciprocity," *J. Theor. Biol.*, vol. 194, pp. 561–574, 1998.
- [21] —, "Evolution of indirect reciprocity by image scoring," *Nature*, vol. 393, pp. 573–577, 1998.
- [22] X. Yao and P. Darwen, "How important is your reputation in a multi-agent environment," in *Proc. 1999 Conf. Syst., Man, Cybern.*, 1999, pp. 575–580.
- [23] X. Yao, "Evolving artificial neural networks," *Proc. IEEE*, vol. 87, no. 9, pp. 1423–1447, Sep. 1999.
- [24] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster," *IEEE Trans. Evol. Comput.*, vol. 3, no. 2, pp. 82–102, 1999.
- [25] K. Chellapilla and D. B. Fogel, "Evolution, neural networks, games, and intelligence," *Proc. IEEE*, vol. 87, no. 9, pp. 1471–1496, Sep. 1999.
- [26] G. Roberts and T. N. Sherratt, "Development of cooperative relationships through increasing investment," *Nature*, vol. 394, pp. 175–179, 1998.
- [27] O. Leimar and P. Hammerstein, "Evolution of cooperation through indirect reciprocity," *Proc. Royal Soc. London*, ser. Series B, vol. 268, pp. 745–753, 2001.

Siang Yew Chong (M'99) received the B.Eng. (Hon.) and M.Eng.Sc. degrees in electronics engineering from Multimedia University (MMU), Melaka, Malaysia, in 2002 and 2004, respectively. He is currently working towards the Ph.D. degree in computer science at the University of Birmingham, Birmingham, U.K.

He is a former Tutor for the Faculty of Engineering and Technology, MMU, and is currently a Member of the Natural Computation Research Group, School of Computer Science, University of Birmingham. His

research interests include evolutionary computation, neural networks, and game theory.

Mr. Chong is the recipient of the Student Travel Grant for the 2003 Congress on Evolutionary Computation (CEC'03).

Xin Yao (M'91–SM'96–F'03) received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, in 1982, the M.Sc. degree from the North China Institute of Computing Technology, Beijing, in 1985, and the Ph.D. degree from USTC in 1990.

He was an Associate Lecturer and Lecturer from 1985 to 1990 at USTC, while working towards his Ph.D. He took up a Postdoctoral Fellowship in the Computer Sciences Laboratory, Australian National University (ANU), Canberra, in 1990, and continued

his work on simulated annealing and evolutionary algorithms. He joined the Knowledge-Based Systems Group, CSIRO Division of Building, Construction and Engineering, Melbourne, in 1991, working primarily on an industrial project on automatic inspection of sewage pipes. He returned to Canberra in 1992 to take up a lectureship in the School of Computer Science, University College, University of New South Wales (UNSW), Australian Defence Force Academy (ADFA), where he was later promoted to a Senior Lecturer and Associate Professor. Attracted by the English weather, he moved to the University of Birmingham, Birmingham, U.K., as a Professor of Computer Science in 1999. Currently, he is the Director of the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), a Distinguished Visiting Professor of the University of Science and Technology of China, Hefei, and a visiting professor of three other universities. He has more than 200 publications. He is the Editor-in-Chief of the *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*. He is an associate editor or editorial board member of several journals. He is the Editor of the *World Scientific Book Series on Advances in Natural Computation*. He has given more than 45 invited keynote and plenary speeches at conferences and workshops worldwide. His major research interests include evolutionary artificial neural networks, automatic modularization of machine learning systems, evolutionary optimization, constraint handling techniques, computational time complexity of evolutionary algorithms, coevolution, iterated prisoner's dilemma, data mining, and real-world applications.

Dr. Yao was awarded the President's Award for Outstanding Thesis by the Chinese Academy of Sciences for his Ph.D. work on simulated annealing and evolutionary algorithms. He won the 2001 IEEE Donald G. Fink Prize Paper Award for his work on evolutionary artificial neural networks. He is the Editor-in-Chief of the *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*.