Learning to cooperate in the Prisoner's Dilemma: Robustness of Predictions of an Instance-Based Learning Model

Cleotilde Gonzalez (coty@cmu.edu)
Dynamic Decision Making Laboratory, Social and Decision Sciences Department
Carnegie Mellon University. Pittsburgh PA 15217 USA
Noam Ben-Asher (noamba@cmu.edu)
Dynamic Decision Making Laboratory, Social and Decision Sciences Department
Carnegie Mellon University. Pittsburgh PA 15217 USA

Abstract

The dynamics of cooperation in repeated Prisoner's Dilemma (PD) interactions are captured by an instance-based learning model that assumes *dynamic adjustment* of expected outcomes (IBL-PD model). This research presents this model's predictions across a large number of PD payoff matrices, in the absence of human data. Rapoport and Chammah (1965) test three hypotheses in a large set of PD payoff matrices: (1) as reward of cooperation increases, more cooperation is observed; (2) as the temptation to defect increases, less cooperation is observed; and (3) as punishment for defection increases, more cooperation is observed. We demonstrate that the same IBL-PD model that was found to predict the dynamics of cooperation in one particular payoff matrix of the PD produces accurate predictions of human cooperation behavior in six additional games. We also make detailed predictions of the dynamics of cooperation that support these three hypotheses.

Keywords: Instance-Based Learning Theory; Cognitive Modeling; Prisoner's Dilemma; Cooperation; Trust.

Learning to Cooperate

Productive endeavors in society often rely on people's willingness to engage in cooperation, even in situations that may be individually costly. People engage in cooperative behavior often: long-term marriages, business initiatives, friendships, and international agreements all depend on expectations of reciprocity that often involve a sacrifice from the individual perspective.

In the laboratory, researchers have studied these social dilemmas using economic games, in which incentives depend on the actions of two players (Rapoport, Guyer, & Gordon, 1976). Extensive research in Economics and Psychology has indicated that in one-time interactions, these games may result in very different behavior than in repeated interactions (Camerer, 2003). However, research on learning that can explain the emergence, sustainability, and adaptability of cooperation from repeated two-agent interactions is very scarce (Gonzalez et al., in press). Existing models underplay the role of cognitive processes necessary to recognize, remember, adapt, and respond to one's prior history of interactions with an opponent (Ben-Asher, Dutt, & Gonzalez, 2013; Gonzalez et al., in press).

Recently, novel results were reported regarding how individuals learn while playing a well-known social dilemma (Prisoner's Dilemma) repeatedly in succession, given different levels of interdependency information (Martin, Gonzalez, Juvina, & Lebiere, in press). Their findings include a significant effect of the information available on the average levels of individual and mutual cooperation, as well as on the trends of cooperation over repeated trials. They found greater cooperation with higher levels of information, and *increasing* trends on the cooperation over trials when fully descriptive information of the payoff matrix was available to the participants.

A large number of models in the BGT tradition make one common assumption: that players have full and complete information regarding the state of the environment, including the actions taken and the outcomes received by themselves and their opponents, as well as each player's forgone payoffs (outcomes that would have been received had one chosen the other option). Also, most of these models make cognitively implausible assumptions, such as players being able to remember large amounts of information. Gonzalez and colleagues (Gonzalez et al., in press) proposed a cognitive model for repeated social interactions that does not require on the full information assumption. In fact, this model, IBL-PD, builds on a model of individual learning and decisions from experience in repeated binary choice (IBL model, Gonzalez & Dutt, 2011; Lejarraga et al., 2012).

The IBL-PD model relies on a formalization of the connection between Social Value Orientation (SVO) research in social dilemmas (Murphy & Ackerman, 2014) and a cognitive learning theory of dynamic decision making (IBLT) (Gonzalez, Lerch, & Lebiere, 2003). By systematically testing a set of assumptions regarding how humans account for the opponent's actions and outcomes into their own decisions, Gonzalez and colleagues conclude that the best IBL-PD model is one that assumes *dynamic adjustment* of expected outcomes rather than static functions of how a player accounts for the opponent's information in evaluating her own actions. This model indicates that the social mechanism that best captures the dynamics of cooperation is the adaptation of dynamic expectations, where players adjust the weight given to the opponent's outcomes based on expectations and the opponent's actual behavior (i.e., surprise).

This model was able to account for several patterns of dynamics in the human data. For example, the fact that even with full information, participants displayed an initial attempt to act selfishly and a gradual discovery of the benefits of cooperation with experience with the same partner.

Although Gonzalez et al. (in press) already demonstrated a number of generalizations of a model that was initially developed to account for individual dynamic decision making, their demonstrations of the IBL-PD model used one single PD payoff matrix. The question we address here is that of the robustness of the predictions that the IBL-PD model can make across a wider range of payoff matrices in the PD. In the current research we produce predictions of average cooperation as well as dynamics of cooperation **in the absence of human data**. Thus, the current work is not a model comparison exercise.

The same IBL-PD model developed by Gonzalez et al. (in press) is used here to test three hypotheses regarding the effects of rewards, temptation to defect, and punishment on the overall proportion of cooperation, proposed by Rapoport and Chammah (1965) (*RC65* hereafter). In addition to making behavioral predictions on the average proportion of cooperation in seven different matrices (reported in RC65), we also make predictions of the dynamics of cooperation in repeated interactions on these seven different payoff matrices (for which human patterns are not reported in RC65). Implications of these results for our understanding of cooperation and trust, and the power and limitations of the IBL-PD model's predictions are discussed.

The Prisoner's Dilemma and Rapoport and Chammah's (1965) hypotheses

The Prisoner's Dilemma (PD) is perhaps the most well-known example of a social dilemma. This is a non-cooperative game in which two players independently choose between two possible strategies: to cooperate or to defect. In the generic game matrix that is illustrated in Figure 1, R refers to the payoff that each player receives as a reward when both players cooperate; S refers to the payoff received by the player who cooperated while the other defected; T refers to the payoff that a player hopes to get if he can defect and get away with it; and P is the payoff of both players when both defect.

		Player 2 Options	
		C	D
Player 1 Options	C	R, R	S, T
	D	T, S	P, P

Figure 1. Generic definition of payoffs in a 2×2 game.

The PD presents a situation with the following rank order of outcomes: S < P < R < T. RC65 defined seven variants of the PD designed to investigate whether the motivations from the relationships of payoffs would be reflected directly in human performance. More precisely, they defined and tested three hypotheses:

1) As R increases, more cooperation will be observed (all else held constant).

2) As T increases, less cooperation will be observed.

3) As P decreases, more cooperation will be observed (all else held constant).

The seven payoff matrices (games) used are shown in Figure 2.

Game I

	C	D
C	9, 9	-10, 10
D	10, -10	-1, -1

Game II

	C	D
C	1, 1	-10, 10
D	10, -10	-9, -9

Game III

	C	D
C	1, 1	-10, 10
D	10, -10	-1, -1

Game IV

	C	D
C	1, 1	-2, 2
D	2, -2	-1, -1

Game V

	C	D
C	1, 1	-50, 50
D	50, -50	-1, -1

Game XI

	C	D
C	5, 5	-10, 10
D	10, -10	-1, -1

Game XII

	C	D
C	1, 1	-10, 10
D	10, -10	-5, -5

Figure 2. Seven payoff matrices in RC65 for the PD

Games III, XI, and I were used to test Hypothesis 1: R increased from 1 to 5 to 9, while T, S, and P were held constant. Games IV, III, and V were used to test Hypothesis 2, where T and S increased from 2 (-2) to 10 (-10) to 50 (-50), while R went from 9 in game IV to 1 in games III and V, and P are held constant. Please note that, this hypothesis suggest that the temptation to defect T is equivalent to the absolute value of the payoff received by the player who cooperated while the other defected. Finally, games III, XII, and II were used to test Hypothesis 3, where P value decreased from -1 to -5 to -9, while R, S, and T were kept constant.

In each interaction (trial), two players repeatedly decided to cooperate or to defect, without communicating and while explicitly given information about the payoff matrix (see Appendix I in RC65 for specific instructions). Participants interacted with the same opponent across 300 trials of one specific game, without being aware of the number of interactions. Their aggregated results (pp. 47, RC65) serve as a benchmark to compare IBL-PD model's predictions.

The IBL-PD model

The IBL-PD model reported in Gonzalez et al. (in press) is an extension of a cognitive model of individual repeated binary choice (Lejarraga, Dutt & Gonzalez, 2012). The IBL-PD includes two IBL models representing individuals making selections between two options, and they are "connected" to one another according to the PD's actions and outcomes in a payoff matrix. The IBL-PD model relies on the concept of Social Value Orientation (SVO) to test hypotheses regarding how a player would account for the opponent's outcome when evaluating one's own potential actions.

The IBL model of individual binary choice has been described in detail in many publications (e.g., Gonzalez & Dutt, 2011; Lejarraga, et al., 2012), and it can be summarized as follows:

In each trial t, the IBL model chooses an option with the highest *Blended Value* (V_j). The V of j is evaluated as:

$$V_j = \sum_{i=1}^{n} p_{ij} x_{ij} \qquad (1)$$

where x_i is the value of the observed outcome i, and p_i is the probability of retrieving that outcome from memory. At trial t, the *retrieval probability* of the observed outcome i is a function of that outcome's activation relative to the activation of all the observed k outcomes for option j.

$$P_{ij} = \frac{e^{\frac{A_i}{\tau}}}{\sum_j e^{\frac{A_j}{\tau}}} \qquad (2)$$

where τ is random noise defined as $\tau = \sigma\sqrt{2}$, and σ is a parameter fitted to the data (see below). At trial t, the activation (Anderson & Lebiere, 1998) of an outcome i is:

$$A_i = ln\left(\sum_{t_i \in \{1,\dots,t-1\}} (t - t_i)^{-d}\right) + \sigma \cdot ln\left(\frac{1-\gamma_i}{\gamma_i}\right) \qquad (3)$$

where d is a decay parameter fitted to human data (see below); $\gamma_{i,t}$ is a random draw from a uniform distribution bounded between 0 and 1 for each outcome and trial; and t_i is each of the previous trial indexes in which the outcome i was encountered.

Two main extensions to account for the dynamics of cooperation in PD (Gonzalez et al., in press) are:

1) Integration of Social Value Orientation in the Blending Equation.

Social Value Orientation (SVO), defined as the degree to which somebody cares about the outcomes of others (Balliet, Parks, & Joireman, 2009), was integrated into the IBL's blending mechanism by replacing Equation 1 with the following blending equation:

$$V_j = \sum_{i=1}^{n} p_{ij} \left(x_{ij} + w\, o_{ij}\right) \qquad (4)$$

Where x_{ij} and o_{ij} are the values of the player's outcome and the opponent's outcome, respectively, in instance i associated with the option j; w represents the extent to which a player considers the other player's outcome for each option when attempting to make a choice that maximizes gains in each trial; and p_{ij} is the probability of retrieving the instance i associated with option j from memory (see Equation 2). Note that o_{ij} represents the opponent's outcome (which comes from its choice of option C or D), given that the player chooses option j.

2) Dynamic expectations and choice as a function of surprise.

Gonzalez et al. (in press) proposed a dynamic adjustment of w as a function of the gap between expected outcomes and the actual outcomes obtained in a trial. The difference between the expected outcomes and those actually obtained are referred to as a "surprise," (disconfirmed expectations) (Maguire, Maguire, & Keane, 2011). In this model, w_t (the regards to the opponent's outcome at trial t) was defined as follows:

$$w_t = 1 - Surprise_t \qquad (7)$$

where w_t varies between selfish behavior w_t=0 and complete fairness w_t=1 according to the surprise resulting from the difference between expected and actual opponent's behavior in that trial. Surprise at trial t, was defined according to past formulations (Erev, Ert, & Roth, 2010; Gonzalez, Dutt, & Lejarraga, 2011) as follows:

$$Surprise_t = \frac{Gap_t}{[Mean(Gap_t) + Gap_t]} \qquad (8)$$

Where the *Gap* at time t is the absolute value of the difference between the expected utility for choosing option j, which is V_j at the previous time period (t-1), and the corresponding actual joint outcome, which is the sum of the player's outcome and the opponent's outcome: (X_{ij} + O_{ij}). Defined as follows:

$$Gap_t = Abs\left[V_{j(t-1)} - \left(X_{ij} + O_{ij}\right)\right] \qquad (9)$$

The *Mean Gap* at time t was defined assuming a horizon of 200 trials of repeated PD as follows:

$$Mean\ (Gap_t) = Mean\,(Gap_{t-1})\left(1 - \frac{1}{200}\right) + Gap\,(t)\left(\frac{1}{200}\right) \qquad (10)$$

IBL-PD model's predictions and the effects of payoffs on Cooperation

Gonzalez et al. (in press) fitted the IBL-PD model to data reported in Martin et al. (in press), which was collected using Game III (Figure 2). The fitting of the d and σ parameters in the IBL-PD model led to values of 2.038 and 0.495, respectively

For the current research we generate predictions in the absence of human data in the additional matrices shown in Figure 2. We ran this model for 100 simulated pairs of players in each of these matrices and produced data to test the different hypotheses in RC65. The human data presented in the following figures come from our estimation of the average values found in the graphs published in RC65. We do not possess any human data to demonstrate those effects.

Hypothesis 1: As reward (R) increases, more cooperation is observed (all else being constant).

Figure 3 shows the average proportion of cooperation resulting from the IBL-PD model's predictions and the proportion of cooperation reported in the RC65 experiments, to test Hypothesis 1. In agreement with the RC65 results, the IBL-PD model predicts that as the reward for cooperation (CC cell) goes from 1 to 5 to 9 (while all other payoffs stay the same), the proportion of cooperation increases.

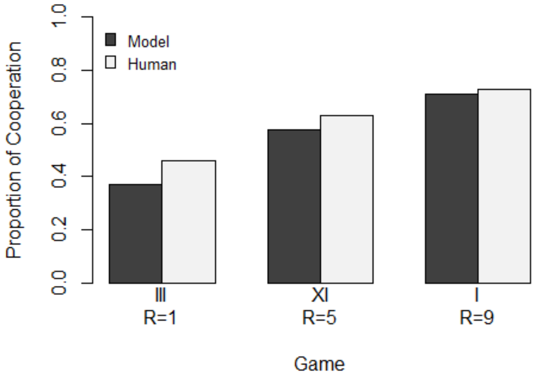

Figure 3. Average proportion of cooperation resulting from the human data reported in RC65 and the IBL-PD model predictions in matrices III, XI, and I: showing that as the mutual cooperation reward increases, the proportion of cooperation increases.

Hypothesis 2: As temptation (T) increases, less cooperation is observed.

Figure 4 shows the average proportion of cooperation observed in the RC65 human experiments and the IBL-PD model's predictions for Hypothesis 2. RC65's data corroborate their hypothesis showing a decrease in cooperation as the temptation to defect, T increases from 2 to 10 to 50. The IBL-PD model shows a decrease in cooperation from Game IV to Game III, but not a decrease from Game III to Game V. The higher cooperation in Game IV compared to the other games may be explained by the relative small advantage for cooperation over defection in this game (T-R=1), compared to the Games III and V where the gain from defection is much larger compared with the gain from cooperation (T-R=9 and 49, respectively). Thus, and as we will observe in the next section, learning about the benefit of cooperation require longer repeated interaction. As the value of *w* (Equation 4) gets closer to 1 with repeated experiences, the model considers the opponent's outcome as equally important as the player's own outcome. Given the exact asymmetry of S and T and the RC65 assumption that T=|S| that we also adopted in the model, S and T values "cancel" each other out as *w* is closer to 1, making the model less sensitive to the mixed CD, DC actions. As a result the model becomes more sensitive to the reward (CC action), and given that R is the same in both, Game III and Game IV, the model ends up making similar overall proportion of cooperation in these two games.

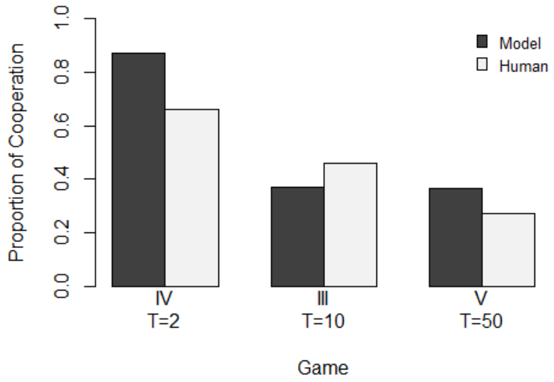

Figure 4. Average proportion of cooperation resulting from the human data reported in RC65 and the IBL-PD model predictions in matrices IV, III, and V: showing that as the temptation to defect increases, the proportion of cooperation decreases.

Hypothesis 3: As punishment (P) increases, more cooperation is observed (all else being constant).

Figure 5 shows the average proportion of cooperation observed in RC65 human experiments and the IBL-PD model's predictions for Hypothesis 3. RC65's data show that as the outcome from mutual defection (DD cell) increases from 1 to 5 to 9 (all else held constant), the proportion of cooperation increases. The IBL-PD model makes accurate predictions of this trend.

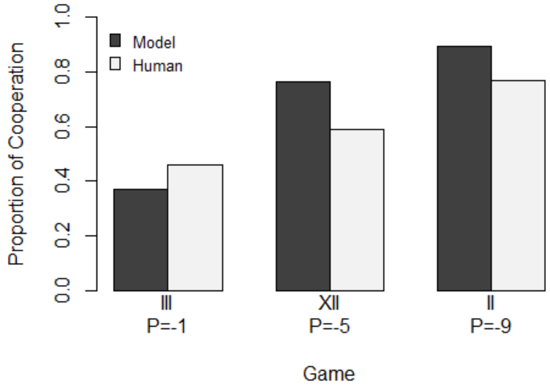

Figure 5. Average proportion of cooperation resulting from the human data reported in RC65 and the IBL-PD model predictions in matrices III, XII, and II: showing that as the punishment for defection increases, the proportion of cooperation increases.

IBL model's predictions of dynamics of cooperation

We make predictions of the dynamics of cooperation over 200 trials that correspond to the three hypotheses of RC65.

As seen in Figure 6, which corresponds to Hypothesis 1, the trends in the proportion of cooperation over time are sensitive to the value of the reward from mutual cooperation. As the magnitude of the reward increases, the IBL-PD model converges sooner towards increasing cooperation, through learning that cooperation is more beneficial when the reward is larger (R=1, 5, 9). Importantly, and as observed in Gonzalez et al. (in press), the dynamics of cooperation in the short-term are different from those in the long-term. In all the games, there is an initial tendency towards *decreased* cooperation. This initial dent lasts longer when the reward is small (Game III) than when the reward in larger (Games XI, and I). The magnitude of the reward influences the initial dent of cooperation and the speed of the increasing trend towards cooperation.

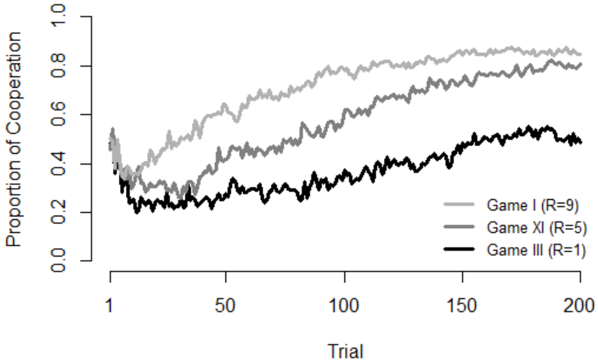

Figure 6. Dynamics of cooperation as a function of the increase in the reward for mutual cooperation.

Although from the perspective of reward, Game XI is halfway between Games III and I, it seems that the proportion of cooperation in Game XI at trial 200 is more similar to Game I than to Game III. The learning process in Game III seems to be slower, probably due to the relatively larger difference between R and T (R-T=9) in this game, compared to the two other games (4 and 1, respectively).

Figure 7 illustrates the dynamics of cooperation as the temptation to defect increases that correspond to the Hypothesis 2. As observed, dynamic behavior in game IV is clearly different from behavior in games III and V. In Game IV the initial decrease of cooperation does not occur. Rather soon, participants realize that the temptation to defect (T=2) is not worth when compared to the reward for cooperation in this game (R=1). For this game, the model quickly learns that the most beneficial decision is to cooperate; and the proportion of cooperation reaches about 80% in less than 50 trials, then ultimately to about 100% within 100 trials. In contrast, the higher temptation values in Games III and V (T=10 and 50, respectively), compared to the reward from cooperation in both of these games (R=1), leads to an initial decrease in the proportion of cooperation. This initial dent is deeper in Game V than in Game III. However, as observed in the dynamics, the increase towards cooperation after the initial decrease occurs faster for Game V than for Game III. Eventually by trial 200, the proportion of cooperation in Game V is higher than in Game III. As suggested before, the increase in T from 10 to 50, produces an initial larger tendency to defect in Game V than in Game III. But, as the simulated players start to mutually cooperate and the weight given to the opponent's outcome increases (*w* approaches 1), the value of the mixed actions (CD and DC) starts to decrease (to zero) given the exact asymmetry of the outcomes (e.g. T=+10, S=-10 in Game III) and model's Blending Value formulation (Equation 4).

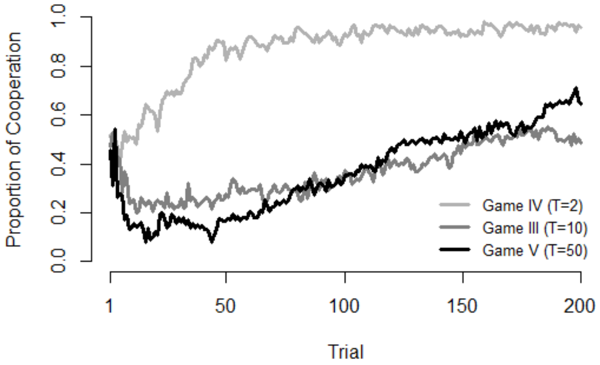

Figure 7. Dynamics of cooperation as a function of the increase in the temptation to defect.

Figure 8 illustrates the dynamics of cooperation as the punishment (P) to defect increases. As the punishment to defect goes from -1 to -5 to -9 with all else being equal, the convergence towards cooperation occurs sooner and faster. High punishment, as in Game II, accelerates learning and shows an immediate rapid increase in cooperation converging into full cooperation rather quickly. High punishment given mutual defection seems to resolve the social dilemma relatively quickly, leading to a relatively stable 100% cooperation after the 50th trial. This behaviour might be attributed to the SVO-inpired consideration of the opponent's payoffs. Considering the other's payoffs amplifies the punishment, as both players receive the same negative payoff for mutual defections. Because the temptation to defect is the same for all of these games, lowering the punishment leads to an initial decrease in cooperation in games XI and III. With low punishment of -1 like in Game III and a relatively high temptation to defect (10), the attempts to benefit from defection seems to last for a much longer compared to other Games.

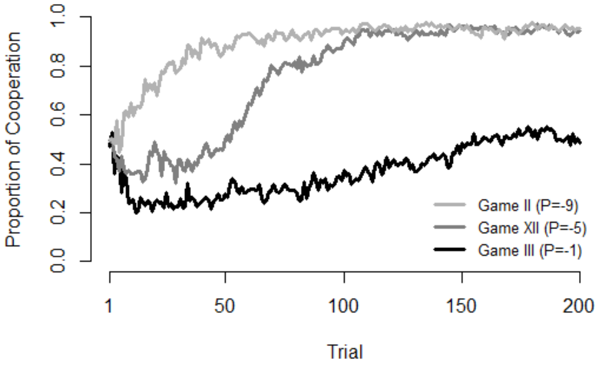

Figure 8. Dynamics of cooperation as a function of the increase in the punishment for mutual defection.

Conclusions

This paper presents novel predictions from a model recently proposed to account for the dynamics of cooperation in repeated social dilemma interactions, IBL-PD (Gonzalez et al., in press). The IBL-PD model is an extension of the individual version of the model (Lejarraga et al., 2012; Gonzalez & Dutt, 2011). In the IBL-PD the concept of SVO is used to test a set of hypotheses about how players in a PD social dilemma consider their opponents' outcomes. In Gonzalez et al. (in press), the model that best accounts for the dynamics of cooperation is one that adjusts the consideration for the opponent's outcome in the player's own decisions as a function of the "surprise": the difference between expectations and actual outcomes.

In the current research we present one important demonstration of the robustness of the IBL-PD model. We use the IBL-PD model without modifications, to make predictions in six additional payoff matrices, in the absence of human data. Data produced from the IBL-PD model reproduces the average behavioral results from RC65 on the proportion of cooperation quite well. Furthermore we report predictions of the dynamics of cooperation in all these additional payoff matrices.

Models that can make general predictions of behavior in multiple tasks are rare (see discussion of this argument in Lejarraga et al., 2013; Gonzalez & Dutt, 2011). The IBL general decision process that makes use of memory and activation mechanisms in the ACT-R theory of cognition (Anderson & Lebiere, 1998) has demonstrated robustness across a large number of tasks. The present work provides predictions of an extension of this model to social dilemmas. The model captures essential phenomena across multiple payoff matrices in the well-known PD.

Future research should compare other models to the IBL-PD model in their ability to make accurate predictions across many payoff matrices in the absence of human data.

Acknowledgments

This research was supported by the U.S. Army Research Laboratory under Network Science-Collaborative Technologies Alliance, Cooperative Agreement No. W911NF-09-2-0053, awarded to Cleotilde Gonzalez. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We thank Hau-yu Wong, Dynamic Decision Making Laboratory, for her review and comments.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations, 12*(4), 533-547.

Ben-Asher, N., Dutt, V., & Gonzalez, C. (2013). Accounting for integration of descriptive and experiential information in a repeated prisoner's dilemma using an instance-based learning model. In B. Kennedy, D. Reitter & R. St. Amant (Eds.), *Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling and Simulation*. Ottawa, Canada: BRIMS Society.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton, NJ: Princeton University Press.

Erev, I., Ert, E., & Roth, A. E. (2010). A choice prediction competition for market entry games: An introduction. *Games, 1*, 117-136.

Gonzalez, C., Ben-Asher, N., Martin, J. M., & Dutt, V. (in press). Emergence of cooperation with increased information: Explaining the process with instance-based learning models. *Cognitive Science.*

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating decisions from experience in sampling and repeated choice paradigms. *Psychological Review, 118*(4), 523-551.

Gonzalez, C., Dutt, V., & Lejarraga, T. (2011). A loser can be a winner: Comparison of two instance-based learning models in a market entry competition. *Games, 2*(1), 136-162.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591-635.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 25*(2), 143-153.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*(1), 176-186.

Martin, J. M., Gonzalez, C., Juvina, I., & Lebiere, C. (in press). A description-experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making.*

Murphy, R. O., & Ackermann, K. A. (2014). Social Value Orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review, 18*(1), 13-41.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma: A study in conflict and cooperation.* Ann Arbor: University of Michigan Press.

Rapoport, A., Guyer, M. J., & Gordon, D. G. (1976). *The 2X2 game.* Ann Arbor: University of Michigan Press.