Game-Theoretic Robustness in Cooperation and Prejudice Reduction: A Graphic Measure

Patrick Grim[1], Randy Au[2], Nancy Louie[3], Robert Rosenberger[1], William Braynen[4], Evan Selinger[5], and Robb E. Eason[1]
Group for Logic and Formal Semantics

[1] Department of Philosophy, SUNY at Stony Brook, Stony Brook NY 11794-3750
[2] Department of Communication, Cornell University, Ithaca, NY 14863
[3] Department of Linguistics, Univ. of Southern California, Los Angeles, CA 90007
[4] Department of Philosophy, Arizona State University, Tucson AZ 85721
[5]Department of Philosophy, Rochester Institute of Technology, Rochester, NY 14623-5604
pgrim@notes.cc.sunysb.edu

Abstract

Talk of 'robustness' remains vague, despite the fact that it is clearly an important parameter in evaluating models in general and game-theoretic results in particular. Here we want to make it a bit less vague by offering a graphic measure for a particular kind of robustness—'matrix robustness'—using a three-dimensional display of the universe of 2 x 2 game theory. In a display of this form, familiar games such as the Prisoner's Dilemma, Stag Hunt, Chicken and Deadlock appear as volumes, making comparison easy regarding the extent of different game-theoretic effects. We illustrate such a comparison in robustness between the triumph of Tit for Tat in a spatialized environment (Grim 1995, Grim, Mar, and St. Denis 1998) and a spatialized modeling of the Contact Hypothesis regarding prejudice reduction (Grim, et. al 2005a, 2005b). The geometrical representation of relative robustness also offers a possibility for links between geometrical theorems and results regarding robustness in game theory.

Robustness in Game Theory

Though robustness has long been recognized as an important parameter for evaluating game-theoretic results, talk of 'robustness' generally remains vague.

Tit for Tat (TFT) is widely respected as a 'robust' strategy in the iterated Prisoner's Dilemma. TFT appears as the winner among significantly different groups of submitted strategies in Robert Axelrod's two round-robin computer tournaments. It appears again as the winner in the significantly different biological replication model constructed by Axelrod and William Hamilton (Axelrod 1984). TFT is the winner yet again in a spatialized cellular automata instantiation of the iterated Prisoner's Dilemma using the basic reactive strategies (Grim 1995; Grim, Mar & St. Denis 1998). Axelrod asks "…does [TFT] do well in a wide variety of environments? That is to say, is it *robust*?" (Axelrod 1984, 48). These results seem to indicate that the answer is 'yes'.

TFT's success in a wide variety of different models raises one's confidence that TFT is tagging something important for a wide range of competitive interactions. The question of *how* robust this history shows TFT to be, however, has no precise answer, nor does such a history offer any precise way of comparing the robustness of the TFT effect with others.

Our attempt is to make at least some talk of robustness graphic and more precise. We introduce a formal measure for robustness across one of the standard parameters in game theory: the payoff matrix. This does not and cannot offer a measure of robustness for *all* aspects of interest—robustness across differences in updating algorithms, for example. What the measure does show, however, graphically and immediately, is the comparative robustness of game-theoretic effects across changes in payoff matrix.

In recent work, Robert Axelrod and Ross Hammond demonstrate robustness of a game-theoretic result regarding ethnocentrism by showing that the result remains when important parameters of the model are either doubled or halved (Axelrod & Hammond 2003, 13). We applaud this as a move in precisely the right direction: toward a more formal measure of an intuitively important evaluational criterion for models. Unfortunately, however, the specific 'doubling and halving' measure that Axelrod and Hammond propose is sensitive to the initial choice of parameters. A measure designed to assure us that a result is robust is itself still fragile with respect to the base model chosen.

The approach we outline here removes this difficulty, at least for the parameter of payoff matrix, by offering a standard measure of robustness in terms of the universe of game theory as a whole.[1]

The Cube Universe of 2 x 2 Game Theory

[1] Source code available upon request.

The overwhelming bulk of work in applied game theory is in two-person game theory. The overwhelming bulk of applied work in two-person game theory, moreover, has concentrated on one game in particular: the Prisoner's Dilemma. Over the past 25 years, furthermore, the vast majority of game-theoretic work on cooperation, altruism, and generosity has concentrated on one very particular set of matrix values (or close relatives): the standard matrix for the Prisoner's Dilemma shown in Table 1. Axelrod notes that the two person Prisoner's Dilemma has become "the *E. coli* of social psychology" (Axelrod 1984, 28). It is clear that this particular payoff matrix is the standard laboratory strain.

		Player A: cooperate	Player A: defect
Player B	cooperate	3, 3	0, 5
	defect	5, 0	1, 1

Table 1. Standard Prisoner's Dilemma matrix, left gains to player B, right gains to A

We can find no body of theory that justifies the primary role that these particular values have played. The notion seems widespread, moreover, that results established using just these particular values can be taken as results for the Prisoner's Dilemma in general; only a few pieces of work have explicitly highlighted variance of applicational results across different matrices which fit the requirements of the Prisoner's Dilemma (Nowak & May 1993; Lindgren & Nordahl 1994; Braynen 2004).

Only slightly more justification has been given for obsessive concentration on the Prisoner's Dilemma. William Poundstone writes that "The prisoner's dilemma is apt to turn up anywhere a conflict of interests exists" (Poundstone 1992, 9). Brian Skyrms, on the other hand, has recently argued that exclusive concentration on the Prisoner's Dilemma is a mistake. Skyrms argues that Stag Hunt should be a focal point for social contract theory, particularly with an eye to game dynamics. Many situations that may appear to be Prisoner's Dilemmas, he argues, are rather Stag Hunts in disguise (Skyrms 2004).

The universe of 2 x 2 game theory extends far beyond the particular values of the standard matrix in Figure 1, of course, and far beyond the inequalities definitional of the Prisoner's Dilemma. For different inequalities between our values CC, CD, DC, and DD, we get different games:

DC > CC > DD > CD CC > (CD + DC) /2	Prisoner's Dilemma
DC > DD > CC > CD	Deadlock
DC > CC > CD > DD	Chicken
CC > DC > DD > CD	Stag Hunt

The full universe of 2 x 2 game theory extends beyond these named games as well, including all sets of four possible values for CC, DC, CD, and DD.

The robustness measure we propose consists of a map of this larger universe of game theory. In such a map, the fact that a particular game-theoretic effect holds at a particular set of matrix values can be represented by plotting a particular point in the universe of game theory. One can thus imagine clouds of points representing the various matrices at which a particular game-theoretic effect appears. An effect that is robust across changes in matrix values will occupy a large volume of the game-theoretic universe. A 'fragile' result, on the other hand, will be restricted to particular points or to a small area. Such a map would give us important comparative results as well. One result or effect A could clearly be said to be more robust than another result B if the volume of matrix values for which B holds is included as a sub-volume within the more extensive volume of effect A.

How are we to envisage the universe of 2 x 2 game theory? Because our matrices are written in terms of four basic parameters—CD, CC, DD, and DC—the first inclination is to envisage such a universe as a hyperspace in 4 dimensions. That thought is intimidating, however, simply because of the difficulties of envisaging and conceptually manipulating results in four-dimensional space. What we propose instead is a manageable three-dimensional image of the universe of game theory. The key is that 2 x 2 games are defined in relative rather than absolute terms.

We lose nothing in mapping the universe of game theory if we envisage it in terms of three of our dimensions relative to a fourth. We can, for example, set CC at a constant value of 50 across our comparisons. Values for our variables CD, DC, and DD can be envisaged as values relative to that CC, extending for convenience from 0 to 100. (A complete picture of the universe would extend these values indefinitely in one direction.) Within such a framework, for example, a set of values DC > DD > CC > CD of 5 > 3 > 1 > 0 can be 'normalized' to a CC of 50, giving us 83 1/3 > 50 > 16 2/3 > 0, or approximately 83 > 50 > 17 > 0.

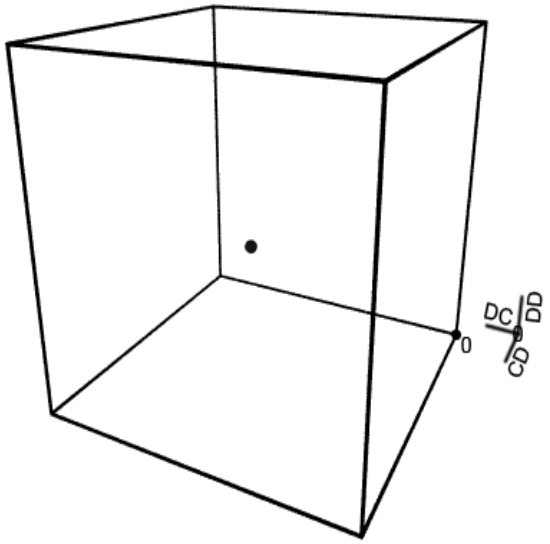

Figure 1. The single most studied point in game theory: The Prisoner's Dilemma with values 5 > 3 > 1 > 0.

Within this universe of game theory, Figure 1 shows the single most studied point: the Prisoner's Dilemma with the standard values of 5 > 3 > 1 > 0. A ball represents this single point. Figure 2 shows the range of the Prisoner's Dilemma, strictly defined with the constraint that CC > [DC + CD] / 2. The volumes corresponding to Stag Hunt, Chicken, and Deadlock are shown in Figures 3, 4, and 5.

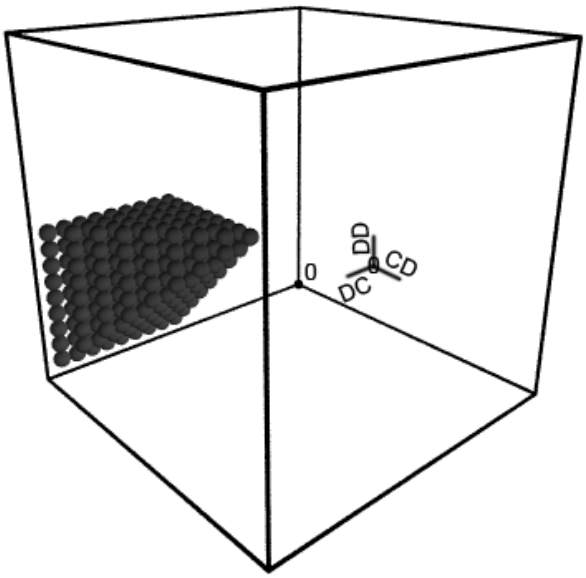

Figure 2. Prisoner's Dilemma

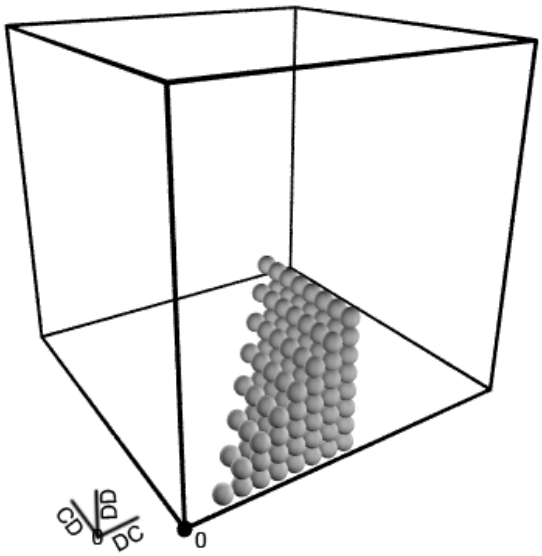

Figure 3. Stag Hunt

Figure 6 shows the complex of these named games as a whole. Here it should be remembered that we are still portraying just a 'chunk' of the game-theoretic universe; a full array would show Deadlock, Chicken, and the Prisoner's Dilemma extending in the direction of the DC and DD axes. Fully rotating versions of these and later illustrations can be found at www.ptft.org/robustAlife.

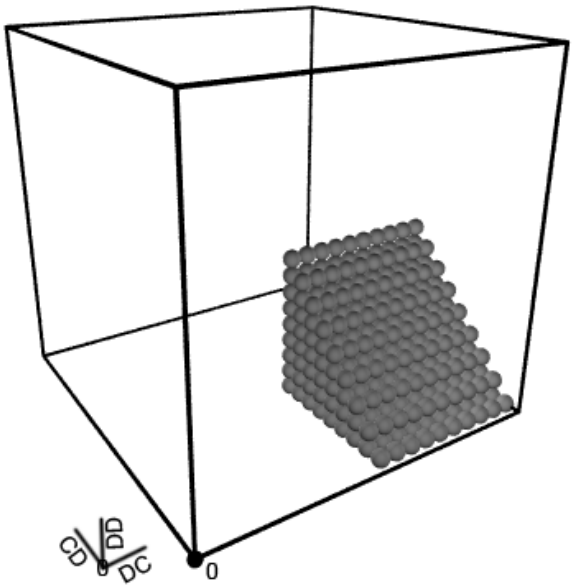

Figure 4. Chicken

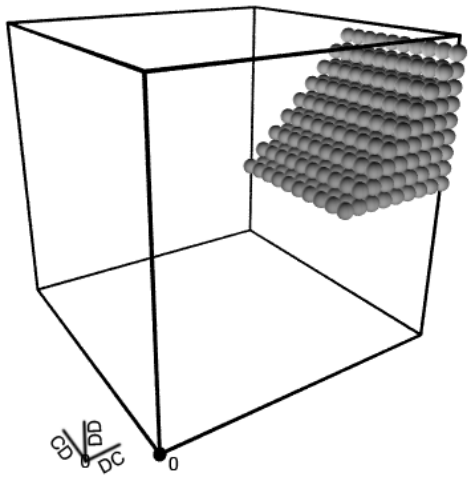

Figure 5. Deadlock

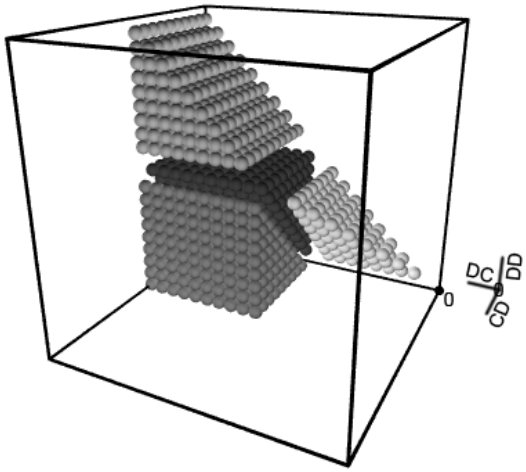

Figure 6. The complex of standard games

The Robustness of TFT

This graphical map of the universe of 2 x 2 game theory offers a measure of robustness across changes in game-theoretic matrices. Points in the graph represent game-theoretic matrices; for a survey of matrix points, we can establish whether a particular game-theoretic result holds at those matrices. Effects which are more robust with respect to matrix changes can generally be expected to be visible across a relatively larger volume of the game-theoretic cube. Comparatively less robust or more fragile effects will be confined to a smaller visible area. Here we offer two examples of the application of the matrix robustness measure.

We will use the term "TFT effect" to refer to TFT's domination over the other seven reactive strategies. This effect, we have noted, has a reputation as a robust effect across different forms of competition. Concentrating on spatialized conquest by TFT in particular, our question will be how robust the spatialized TFT effect is across changes in matrix values.

We use as our basis just the 8 reactive strategies in an iterated Prisoner's Dilemma: those strategies whose behavior on a given round is determined entirely by the behavior of the opponent on the previous round. Using 1 for cooperation and 0 for defection, we can code these 8 basic strategies as 3-tuples <*i*, *c*, *d*>, where *i* indicates a strategy's initial play, *c* its response to cooperation on the other side, and *d* its response to defection:

<0,0,0> All-Defect
<0,0,1> Suspicious Perverse
<0,1,0> Suspicious Tit for Tat
<0,1,1> D-then-All-Cooperate
<1,0,0> C-then-All-Defect
<1,0,1> Perverse
<1,1,0> Tit for Tat
<1,1,1> All-Cooperate

We begin with a randomization of these strategies across a 64x64 cellular automata array.[2] Each cell plays 200 rounds of an iterated Prisoner's Dilemma with its 8 immediate neighbors, then totals its score. If at the end of 200 rounds a cell has a neighbor that has amassed a higher total score, it converts to the strategy of that neighbor. If not, it retains its strategy. Updating is synchronous. In the case of a tie between highest-scoring neighbors, one is chosen at random (Grim, Mar, & St. Denis 1998).

Using the standard DC > CC > DD > CD values of 5 > 3 > 1 > 0 of the Prisoner's Dilemma, it is well known that dominance in such an array goes first to a pair of exploitative strategies: All-Defect (All-D) and C-then-All-Defect (C-then-All-D). Once a range of vulnerable strategies has been eliminated, however, clusters of TFT start to grow, eventually conquering the entire array (Figure 7).

What this shows is spatialized conquest by TFT for the specific matrix values of 5 > 3 > 1 > 0. But how robust is that effect across changes in matrix values?

[2] A change to the eight reactive strategies, or their configuration at the start of play would significantly alter the effect. However, the robustness of effects within these different configurations or populations of strategies could be visualized on our graphic measure.

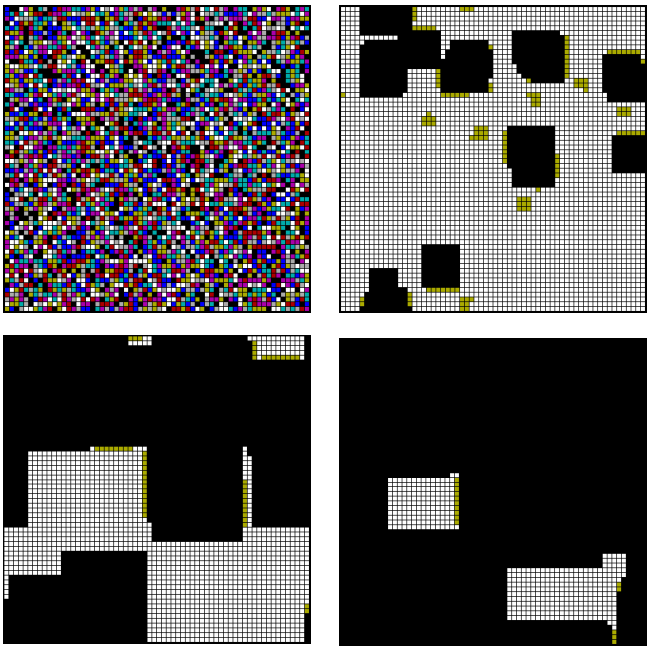

Figure 7. Conquest by TFT in a randomized environment of 8 reactive strategies. TFT is represented in black, All-D in white.

In order to answer that question, we took results across 8,000 spatialized competitions, using values for DD, CD, and DC between 0 and 20 and with CC normalized at a value of 10. In each case we began with a randomization of the 8 reactive strategies across a 64 x 64 array, precisely as above. Those matrix values at which TFT showed a greater than 90% occupation of the array after 100 generations were counted as positive for the TFT effect. Those that showed a lower role for TFT were counted as negative.

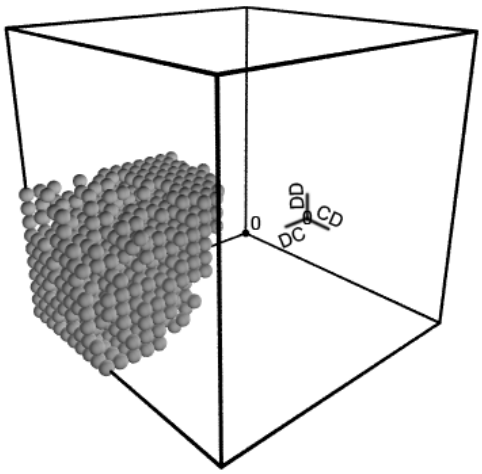

Figure 8. Robustness of the spatialized TFT effect

When plotted, these points give us a clear indication of the robustness of the spatialized TFT effect across changes in matrix values (Figure 8). We take this to be a robust effect, covering area of the Prisoner's Dilemma, into parts of Chicken, and more. A fully rotating image of the result can be found at www.ptft.org/paq/robustAlife.

The Robustness of the Contact Hypothesis

In this section, we offer another effect for comparison: a game-theoretic instantiation of the contact hypothesis.

Despite numerous theories in the social psychological literature regarding the nature and sources of prejudice, there is only one major theory of prejudice reduction: the contact hypothesis. Under the right conditions, the contact hypothesis posits that prejudice between groups will be reduced with increased contact between members of the groups (Allport 1954; Pettigrew 1998).

A computational model for such a hypothesis would need to include at least the following features: (i) distinct groups, (ii) behaviors which may or may not be depend upon the group affiliation of a cell or its neighbor, (iii) advantages and disadvantages resulting from these behaviors, (iv) an updating mechanism for behavior, and (v) configurations of greater and lesser contact.

In earlier work in *Artificial Life IX* we constructed a game-theoretic model of this type for the contact hypothesis (Grim et. al. 2004; 2005). As in the case of the spatialized Prisoner's Dilemma outlined above, cells play only with their eight contiguous neighbors. After 200 rounds of interaction, they adopt the strategy of their most successful neighbor. Although we appropriate the standard payoff matrix and the standard eight reactive strategies, our model is novel in two respects. (1) Each cell is defined not only by strategy, but also by color; each cell is either red or green, and a cell's color never changes during play. (2) One color-sensitive strategy, named Prejudicial Tit for Tat (PTFT), is added to the mix; it plays All Defect against cells of the other color and TFT against cells of its own color.

By varying how the cells are distributed—playing some games in an array that is segregated by color and other games in an array is integrated by color (Figure 9)—we are able to assess the success of PTFT in different environments. The contact hypothesis is tested for a simulational environment by contrasting the success of the prejudicial strategy PTFT in a segregated array with its success in the integrated one.

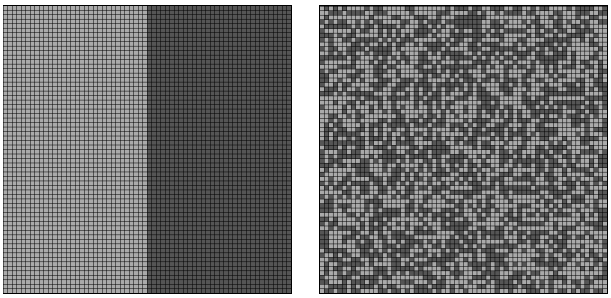

Figure 9. Segregated (left) and mixed patterns of background color

The results show strong computational support for the contact hypothesis. In the segregated array, PTFT and TFT are the only two strategies that remain after approximately 12 generations; each takes up roughly half the area (Figures 10, 11). In the mixed array, on the other hand, TFT eventually takes over nearly the entire array (Figure 12). What the results suggest is that social psychologists should player closer attention to elements of advantage and disadvantage analogous to the game-theoretical mechanisms of such a model.[3]

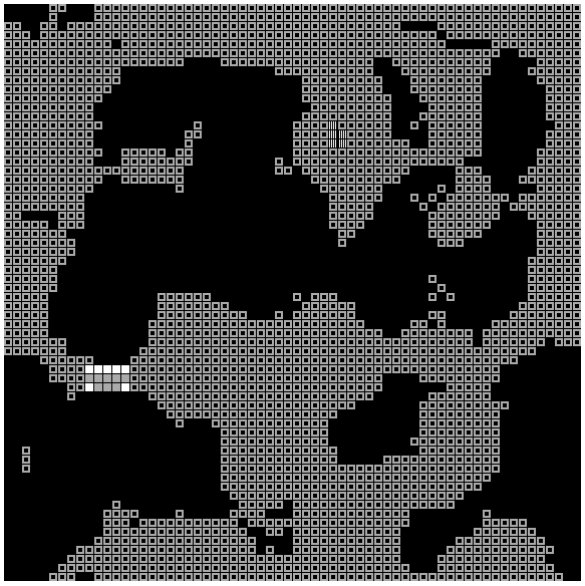

Figure 10. Evolution of randomized strategies to shared dominance by TFT and PTFT in an array segregated by color. A complete evolution can be seen at www.ptft.org/robustAlife.

[3] For more on the philosophical implications of our model for the contact hypothesis see (Grim et al., 2004; 2005).

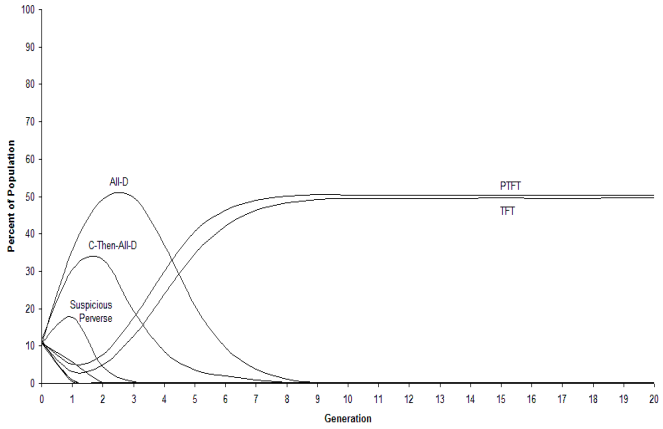

Figure 11. Percentages of the population for 9 strategies in an array segregated by color (20 generations shown). A single, typical run is displayed.

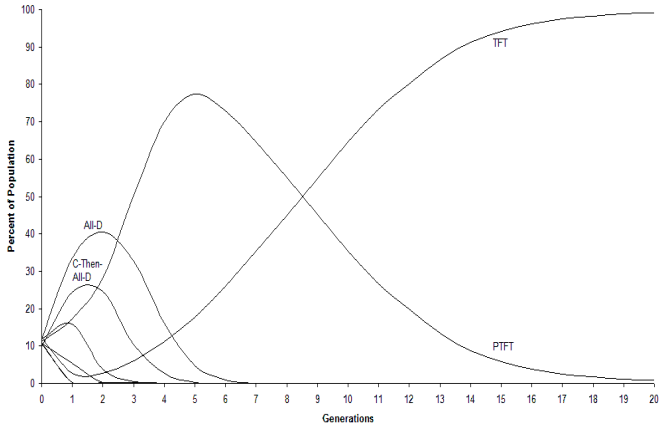

Figure 12. Percentages of the population for 9 strategies in an array randomized by color (20 generations shown). A single, typical run is displayed.

What is at issue here, however, is how robust the PTFT effect is across changes in matrix values. How does it compare, in particular, with the spatialized take-over of TFT in the previous studies?

To investigate which matrices in the game-theoretic universe are ones where the PTFT effect occurs, we plot each point where both TFT takes over more than 90% of the space in a mixed array, and TFT and PTFT each take over more than 40% of the space in a segregated array. Figure 13 shows a graphic portrayal of the matrix robustness of the PTFT effect in these terms.

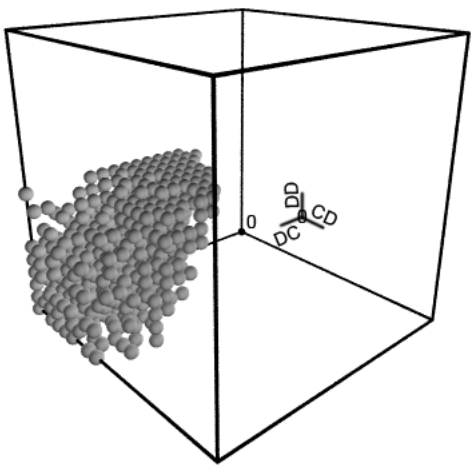

Figure 13. Robustness of the PTFT effect

With two effects in hand, our measure allows a graphic comparison in terms of matrix robustness. TFT, we have noted, is well known as a generally robust strategy. With regard to the specific measure of robustness across changes in matrix values, at least, the PTFT effect outlined here is at least almost as robust as the spatialized TFT effect.

Conclusion

Our attempt here has been to outline and illustrate a new measure for game-theoretic robustness across changes in matrix values.

A single measure adequate for all types of robustness is clearly too much to hope for. What we would like to see is the development of a *number* of standardized measures, adequate for different forms of robustness. Robustness, in all its senses, is a criterion of major importance across modeling quite generally—an importance that underlines the necessity of developing clear measures.

References

Allport, G. W.: (1954), *The Nature of Prejudice*. Addison-Wesley, Cambridge, Mass.

Axelrod, R.: (1984), *The Evolution of Cooperation*, Basic Books, New York.

Axelrod, R. and Hammond, R. A.: (2003), 'The Evolution of Ethnocentric Behavior', Midwest Political Science Convention, April 3-6, Chicago, IL.

Braynen, W.: (2004), *Evolution of Norms and Leviathan*, Master's Thesis, Philosophy, SUNY at Stony Brook.

Grim, P.: (1995), 'The Greater Generosity of the Spatialized Prisoner's Dilemma', *Journal of Theoretical Biology* 173, 353-359.

Grim, P., Mar, G. and St. Denis, P.: (1998), *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. MIT Press, Cambridge, Mass.

Grim, P., Selinger, E., Braynen, W., Rosenberger, R., Au, R., Louie, N. and Connolly, J.: (2004), 'Reducing Prejudice: A Spatialized Game-Theoretic Model for the Contact Hypothesis', in Pollack, J., Bedau, M., Husbands, P., Ikegami, T. and Watson, R. A. (Eds.) *Artificial Life IX*, Cambridge, Mass., MIT Press, pp. 244-249.

Grim, P., Selinger E., Braynen, W., Rosenberger, R., Au, R., Louie, N. and Connolly, J.: (2005), 'Modeling Prejudice Reduction: Spatialized Game Theory and the Contact Hypothesis', *Public Affairs Quarterly* 19, 95-125.

Lindgren, K. and Nordahl, M. G.: (1994), 'Evolutionary Dynamics of Spatial Games', *Physica D* 75, 292-309.

Nowak, M. and May, R.: (1993), 'The Spatial Dimensions of Evolution', *International Journal of Bifurcation and Chaos* 3, 35-78.

Pettigrew, T. F.: (1998), 'Intergroup Contact Theory', *Annual Review of Psychology* 49, 65-85.

Poundstone, W.: (1992), *Prisoner's Dilemma*, Anchor Books, New York.

Skyrms, B.: (2004), *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, New York.