A Graphic Measure for Game-Theoretic Robustness

Patrick Grim, Randy Au, Nancy Louie, Robert Rosenberger, Will Braynen, Evan Selinger, and Robb E. Eason

forthcoming in Synthese

Abstract:

Robustness has long been recognized as an important parameter for evaluating game-theoretic results, but talk of 'robustness' generally remains vague. What we offer here is 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 such a measure specific games appear as specific volumes (Prisoner's Dilemma, Stag Hunt, etc.), allowing a graphic image of the extent of particular game-theoretic effects in terms of those games. The measure also allows for an easy comparison between different effects in terms of matrix robustness. Here we use the measure to compare the robustness of Tit for Tat's well-known success in spatialized games (Axelrod 1984, Grim, Mar, & St. Denis 1998) with the robustness of a recent game-theoretic model of the contact hypothesis regarding prejudice reduction (Grim, Selinger, Braynen, Rosenberger, Au, Louie, & Connolly 2005).

Keywords: robustness, formal measure, game theory, payoff matrix, Prisoner's Dilemma, Tit for Tat, contact hypothesis, social psychology, prejudice, discrimination

I. Model Realism and Robustness

'Robustness' is a primary criterion for evaluating modeling results in general, and game-theoretic results in particular. On seeing a new result, one of the first things a researcher wants to know is how 'robust' that result is--roughly, how well it stands up in variations of the model. The issue at hand is whether very specific choices of parameter values are crucial to the effect, or whether it holds across a broad range of values. Does the result depend on the particular algorithm for reproduction, the specifics of spatial organization or the like, or is it a result that can be expected to appear despite important variations in model structure?

In trying to capture a real phenomenon—physical, chemical, biological, or social—modelers work quite deliberately with a simpler structure. The target reality is often too rich, complex, or messy to be studied directly; the hope is to understand and perhaps predict aspects of that complex reality by working with something that is relevantly analogous but easier to grasp. How well a model captures a target—*how* relevantly analogous any model is—will therefore always be a matter of degree. It will, moreover, remain open for debate whether the model captures the reality well enough—whether it captures essential features rather than inessential details, or deep mechanisms rather than superficial appearances. That crucial question regarding models is one that the models themselves cannot answer.[1]

In building a model, there is always some latitude: investigators may use one core algorithm rather than another, concentrate on one set of parameters rather than others, and standardly test variables within specific ranges. The precise model selected is thus always one out of a range of possible models. If the demonstrated effect shows up only in the specific model selected, one can be no more confident of the reality of the effect than one is confident of the precise accuracy of that model. Given the inherent limitations of modeling, an effect that is 'fragile'—one that is limited to specific choices in a specific model—can therefore be rightly

regarded with suspicion. The touted result may be an artifact of the specific modeling conventions chosen; since one cannot be sure that those are accurate, doubt remains as to whether the effect is real.

'Robust' effects, in contrast, hold for a wide range of models. One reason that robustness is a virtue is that it may raise confidence in the realism of a model. It is generally a better bet that the essentials of a phenomenon will be captured somewhere in a range of possible models than that a single model chosen in that range will happen to capture them precisely. The fact that a phenomenon appears robustly across a range of models therefore increases one's confidence that it is real. It is still possible, of course, that none of the models in the range turns out to be adequate. Robustness rightly builds our confidence in the reality of a modeled phenomenon even though it does not offer any conclusive proof. Conclusive proofs in modeling, as in many aspects of science, are too much to hope for.

What we offer here is a new measure for a particular kind of robustness: a three-dimensional display of the universe of game theory that allows one to compare the prevalence of effects across variations in matrix values (what we will refer to as 'matrix robustness'). We think this constitutes an important measure for game theoretic results, and hope that it will also suggest other measures needed regarding other aspects of robustness.

II. 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.

The history of Tit for Tat (TFT), widely respected as a 'robust' strategy in the iterated Prisoner's Dilemma, serves as a simple example. TFT appears as the winner among significantly different groups of submitted strategies in Robert Axelrod's two round-robin computer tournaments (Axelrod 1980a, 1980b). It appears again as the winner in the significantly different biological replication model constructed by Axelrod and William Hamilton (Axelrod & Hamilton 1981). TFT is once again the winner in a spatialized cellular automata instantiation of the iterated Prisoner's Dilemma using the basic reactive strategies (Grim 1995, 1996; 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 this range of different models raises one's confidence that TFT is tagging something important for a wide range of competitive interactions, both in formal game-theory and in the biological, social, and economic interactions that game-theory is often used to model. But the question of *how* robust this history shows TFT to be has no precise answer, nor does such a history offer any precise way of comparing the robustness of this TFT effect with others.

In what follows, we want to make at least some talk of robustness and of comparative robustness more graphic and more precise. Here 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 reproductive algorithm, for example. What the measure does show, however, graphically and immediately, is 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.

> Not only does ethnocentric behavior evolve in this model, but its emergence is robust under a wide range of parameters. When any of the following parameters are either halved or doubled, at least two-thirds of strategies are ethnocentric and at least two-thirds of the actual choices are ethnocentric: cost of helping, lattice width, number of groups, immigration rate, mutation rate, or duration of the run… (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. The specific 'doubling and halving' measure that Axelrod and Hammond propose, however, is dependant in unfortunate ways on the initial parameters tested. For one set of initial parameters, the 'doubling and halving' measure would vindicate a phenomenon as robust, while for another set it would not. Unfortunately, therefore, the 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. Since that universe of payoff possibilities remains constant, the measure is not sensitive to the particular payoff values with which we first test the phenomenon; it is a measure of robustness that is itself robust. Such an approach, we want to suggest, offers a more objective measure of game-theoretic robustness across changes in payoff matrix and a reliable indicator of the relative robustness of comparative phenomena.

III. The Cube Universe of 2 x 2 Game Theory

The overwhelming bulk of work in game theory to date is work in two-person game theory. Two players are pitted against each other, almost always with just two options of play. What each player gains is dictated by the choices of both players, the results expressed in a 2 x 2 matrix.

Although analytic work in game theory is often more general, the overwhelming bulk of work in applied game theory—game theory applied in simulation to questions of generosity and altruism, for example—has concentrated on one game in particular: the Prisoner's Dilemma. Each player has the option of cooperating or defecting, with payoffs ranked DC > CC > DD > CD. Defection against cooperation (DC) carries a greater payoff for the defector than mutual cooperation (CC), which carries a greater payoff than mutual defection (DD), which carries a greater payoff than cooperating but being defected against (CD). By definition the Prisoner's Dilemma carries a further condition as well; that it not be possible to exceed an average gain of mutual cooperation by alternating defections and cooperations on each side (CC > [DC + CD] / 2).[2]

Over the past 25 years, moreover, the vast majority of game-theoretic simulations regarding cooperation, altruism, and generosity (including our own) have used one particular set of values for the Prisoner's Dilemma, or something close, chosen from the wide universe of 2 x 2

game theory (Axelrod 1980a, 1984, Axelrod and Hamilton 1981, Nowak and Sigmund 1993, Sigmund 1993, Grim 1995, 1996, Wedekind and Milinski 1996, Nakamaru, Matsuda, and Iwasa 1997, Brauchli, Killingback, and Doebeli 1999, Harms 2001, Grim et. al. 2004, 2005). The standard matrix used for the Prisoner's Dilemma is shown below.

		Player B	
		Cooperate	Defect
Player A	Cooperate	**3** , 3	**0** , 5
	Defect	**5** , 0	**1** , 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.

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.[3] William Poundstone writes that "The prisoner's dilemma is apt to turn up anywhere a conflict of interests exists" (Poundstone 1992, 9). Axelrod writes that

> The Prisoner's Dilemma is simply an abstract formulation for some very common and very interesting situations in which what is best for each person individually leads to mutual defection, whereas everyone would have been better off with mutual cooperation. (Axelrod 1984, 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 2001, 2004; see also Bergstrom 2002).

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 > 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, including a full range for values CC, DC, CD, and DD. 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.

What we are proposing is a map of the entire abstract area of 2 x 2 game theory. In some cases, for some questions, nature may dictate a special importance for some sub-region of that space. In that case the techniques we outline could be tailored to that particular issue. Here, however, we concentrate on the general case of the entire abstract space.

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. We routinely exploit the fact that we are evolved from fruit-seeking primates by building on the perceptual abilities that come with that evolutionary history. Visualization in two or three dimensions is of great conceptual benefit, allowing us to tackle formal relations by exploiting immediate perceptual inferences (Larkin & Simon 1987; Grim 2005). But the sad fact is that our spatial abilities are limited to three dimensions. Many of the benefits of visualization are lost if we try to work in four.

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. What qualifies a game as a form of Deadlock, for example, is that DC > DD > CC > CD. Game theory is determined by relative values in a deeper sense as well: the dynamics of a game with values DC > DD > CC > CD of 20 > 10 > 6 > 4 will be identical to a game with values 10 > 5 > 3 > 2. What gives a game its character is not the absolute but the relative values of these variables.

We therefore 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.[4]

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. Of course, the range of the Prisoner's Dilemma is much larger than that point. Figure 2 shows the full range of the

Prisoner's Dilemma, strictly defined with the constraint that CC > [DC + CD] / 2. Figure 3 shows the larger area for a Prisoner's Dilemma in which the additional definitional constraint is dropped. Fully rotating versions of these and later illustrations can be found at www.ptft.org/robustness.

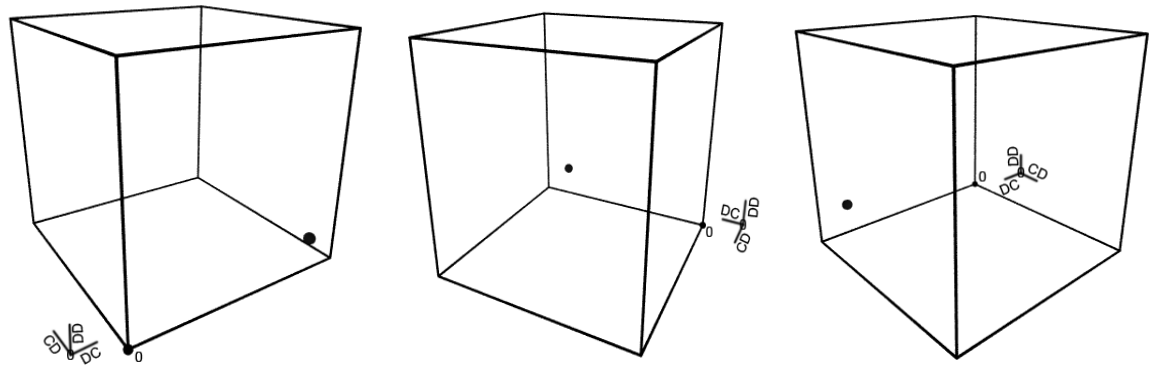

Figure 1. The single most studied point in game theory: The Prisoner's Dilemma with DC > CC > DD > CD values 5 > 3 > 1 > 0

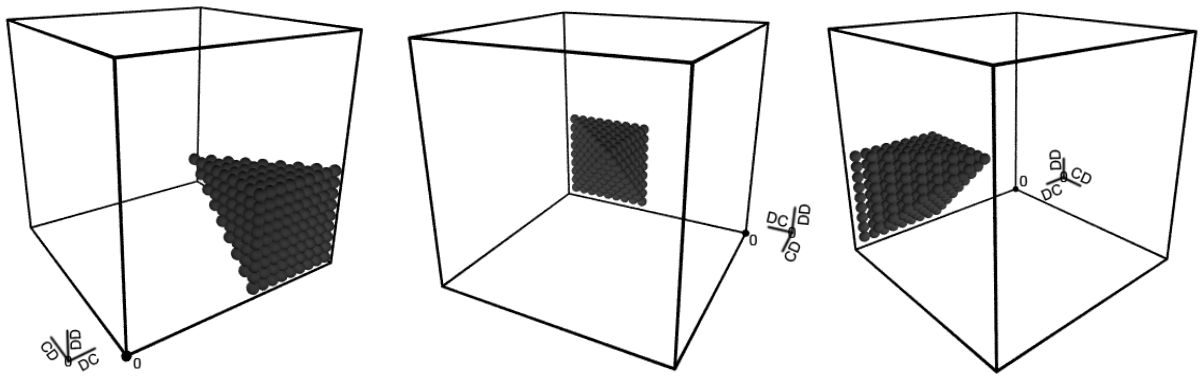

Figure 2. The Prisoner's Dilemma with CC > [DC + CD] / 2

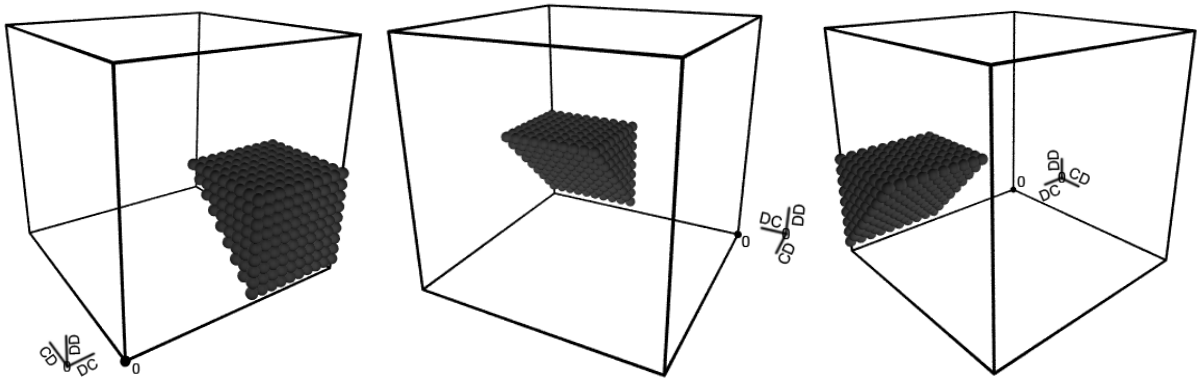

Figure 3. The Prisoner's Dilemma without the standard constraint

The volumes corresponding to Stag Hunt, Chicken, and Deadlock are shown in Figures 4, 5, and 6. In none of Figures 1 through 6 do values go beyond the CD = 50 plane, because we have normalized our cube to CC = 50 and because CD > CC for none of the games defined.

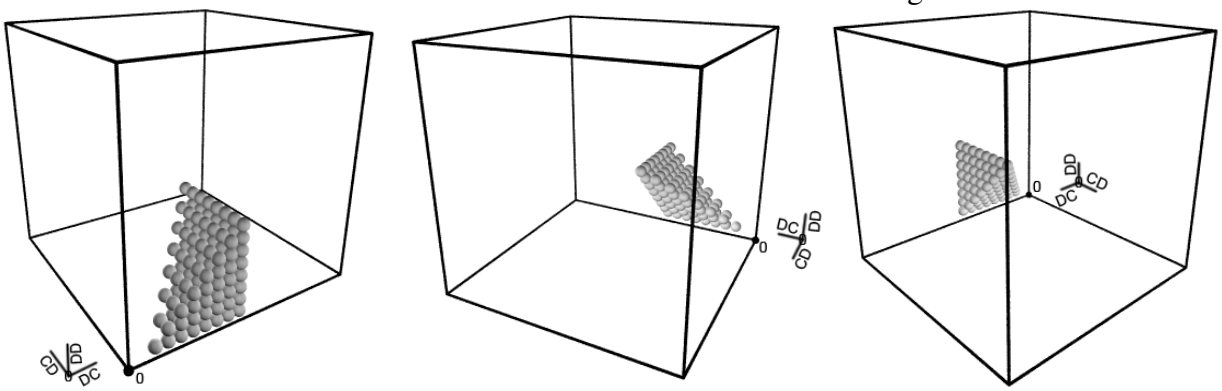

Figure 4. Stag Hunt

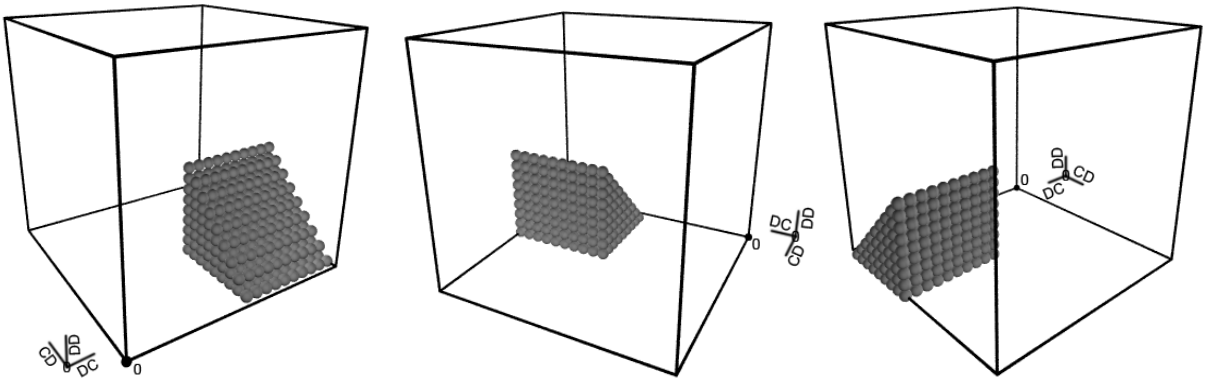

Figure 5. Chicken

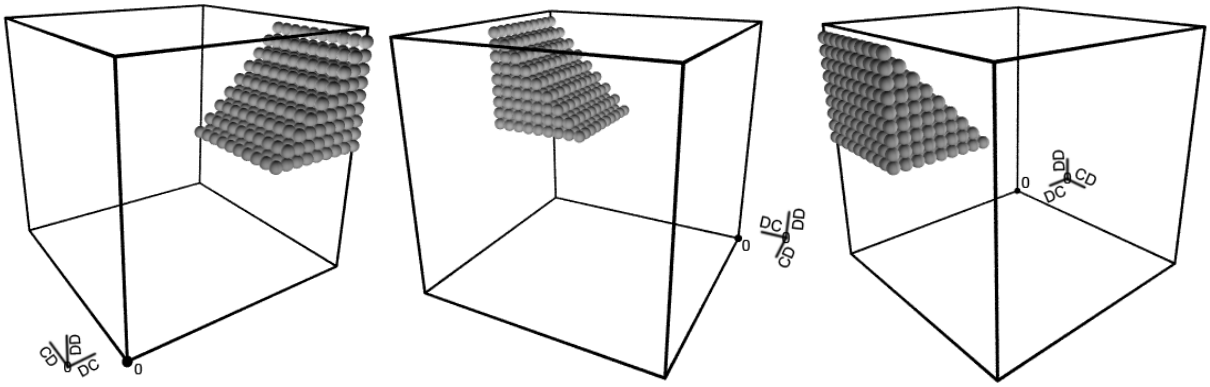

Figure 6. Deadlock

As we have noted, these standard games do not by any means exhaust the universe of game theory. There are 4 factorial or 24 possible inequalities governing our variables CC, CD, DC, and DD, all of which are represented in the universe of game theory but only 4 of which

constitute the games above. The reasons that other games have been ignored are largely interpretational rather than formal. Many have seen cooperation and competition as forming an essential tension in social life. Attention has therefore been concentrated on games in which individual benefit from mutual cooperation conflicts with individual benefit from competition—in which a player's gain from mutual cooperation is greater than from his one-sided cooperation, for example, but in which defection against a cooperator is preferable to mutual defection. Those interests emphasize games in which CC ranks higher than CD and DC higher than DD, and only 6 of the possible orderings satisfy both conditions. Two of those are games in which defection always gets a lower payoff than cooperation, regardless of what the opponent does. If we eliminate those two, we are down to the standard four: the Prisoner's Dilemma without the additional constraint, Stag Hunt, Chicken, and Deadlock (Poundstone 1992). It should be emphasized, however, that what has led us to focus on these games in particular is not merely their formal structure but the informal meanings we give to 'C' and 'D' and our background assumptions about the social and economic life we choose to model.

A significant volume of the game-theoretic cube, comparable to that occupied by these standard games, is occupied by their 'shadows'. Our games are defined in terms of relationships CC, CD, DC, and DD, themselves defined in terms of C and D as options. But what of two games that are symmetrical in the way that the following matrices reflect?

		Player B: Cooperate	Player B: Defect
Player A	Cooperate	**3** , 3	**0** , 5
	Defect	**5** , 0	**1** , 1

		Player B: Cooperate	Player B: Defect
Player A	Cooperate	**1** , 1	**5** , 0
	Defect	**0** , 5	**3** , 3

These two games are different only in that the option called 'defect' in the first game is labeled 'cooperate' in the second. Gains for CC, CD, DC, and DD in the first game are identical to gains

in the second game for DD, DC, CD, and CC; all that has changed is that 'C' appears in place of 'D' and 'D' in place of 'C'. What these matrices represent are thus the Prisoner's Dilemma with the standard values and its 'shadow'.[5]

Although it is not of crucial importance for present purposes, the location of shadow games in the game-theoretic universe is intriguing. If we pile up the game-theoretical volumes for Deadlock, for Chicken, for Stag Hunt, and for the Prisoner's Dilemma without the CC > CD + DC / 2 condition, the mereological whole forms a tight complex on one side of the universe (Figure 7). Here Chicken is a prism lying on the CD-DC floor, Prisoner's Dilemma lies over it, Deadlock sits above the two of them and Stag Hunt is a truncated shape to the right.

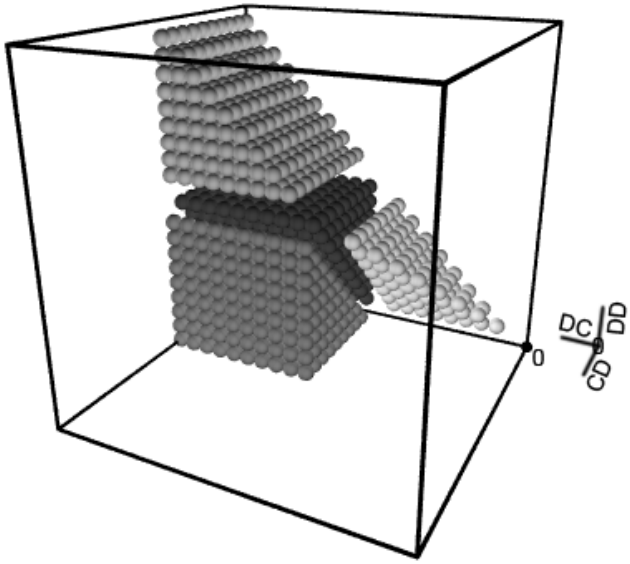

Figure 7. The total game complex
Rotating versions of all illustrations can be found at www.ptft.org/robustness.

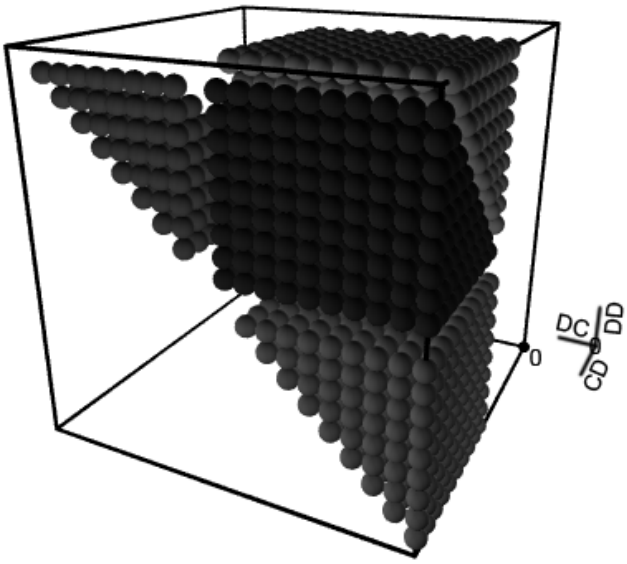

Figure 8. The total shadow complex

The ‘shadow’ of all of these games as a complex lies on the other side of the universe (Figure 8). One way to describe the relative positions of the game complex and its shadow is in terms of new axis labels. The corner farthest from our origin, diametrically opposite across the cube, we might label the ‘counter-origin’. The edge furthest from the DC axis, again diametrically opposite across the cube, we take as our DC’ axis. We also envisage the DC’ axis as running in the opposite direction, starting from 0 at the counter-origin. CD’ and DD’ are similarly the edges furthest from our CD and DD axes, which we envisage as running from 0 at the counter-origin. The relationship between our game complex and its shadow can now be expressed as follows: the shadow complex lies in relation to the counter-origin and axes CD’, DC’, and DD’ precisely as the game complex itself lies in relation to the origin and CD, DC, and DD.

Appropriate rotation of an object in four-dimensional space produces a three-dimensional mirror image of the original (M $\bar{b}$ius 1827 (1976); Rucker 1984). The shadow of our game complex is such a four-dimensional rotation, though also rotated 180E in three dimensions. One could also describe the relative positions of the game complex and its shadow entirely in terms of mirror images. If we take that area corresponding to the game of Deadlock, and take its mirror image across the CD = 50 plane, then take the mirror image of that result across the DC = 50 plane, and finally take the mirror image of that across the DD = 50 plane, we have the position of the Deadlock shadow. The same series of mirror images take us from the game complex as a whole to its shadow as a whole.

What we have tried to describe is the relationship between the game complex as a whole and its shadow as a whole. The same relationship holds for some but not all of its parts and their shadows. The shadow for the Prisoner’s Dilemma without the CC > (CD + DC) / 2 constraint is its 3-way mirror image, as above, as is the shadow for Deadlock (Figures 9, 10). But this does not hold for Stag Hunt and Chicken. In these cases there is a surprising reversal between game and shadow. The shadow for Stag Hunt is the 3-way mirror image not of Stag Hunt but of Chicken (Figure 11). The shadow for Chicken is the 3-way mirror image not of Chicken but of Stag Hunt (Figure 12). Thus, although the game complex and its shadow as a whole stand in the spatial relation outlined, which game occupies which sub-space of that complex changes as we move from the complex to its shadow. Symmetry is also broken between the Prisoner’s Dilemma and its shadow when we include the standard constraint and its appropriate shadow (Figure 13).

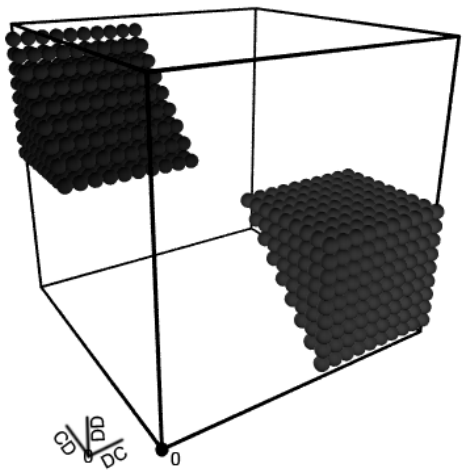

Figure 9. The Prisoner's Dilemma without standard constraint, with shadow

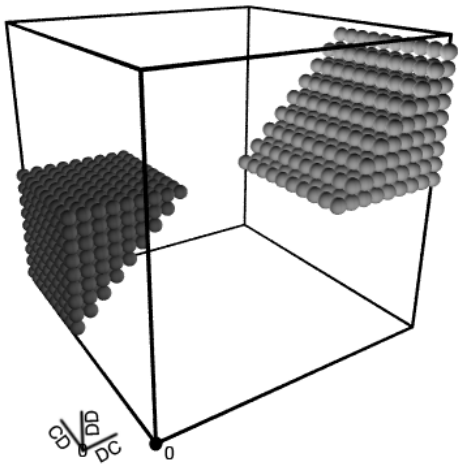

Figure 10. Deadlock, with shadow

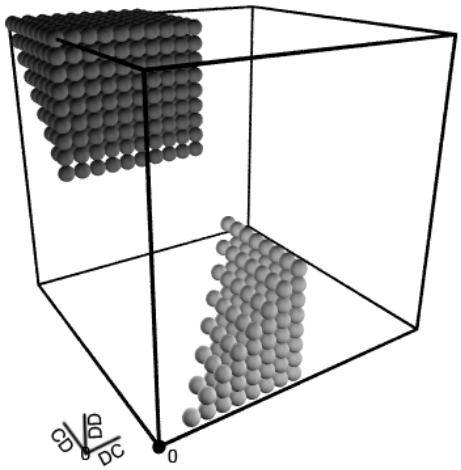

Figure 11. Stag Hunt, with shadow

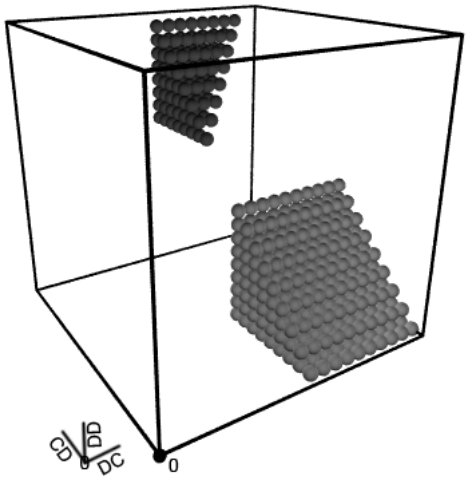

Figure 12. Chicken, with shadow

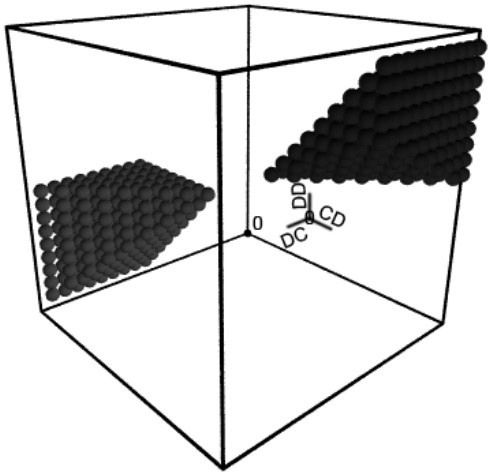

Figure 13. The Prisoner's Dilemma with CC > [CD + DC] / 2 constraint, shadow with symmetrical DD > (CD +DC)/2 constraint

Here and throughout, we deal with a game-theoretic cube in which CC is normalized to 50 and other values are sampled in a range between 0 and 100. Even with CC normalized at 50, of course, the universe of game theory as a whole extends infinitely in the direction of axes CD, DC and DD. Though they capture an important area, therefore, the illustrations above still constitute only a 'chunk' of the whole. As a reminder of this fact, we also offer an illustration with CC normalized at 50 but other values allowed to range between 0 and 200 rather than between 0 and 100 (Figure 14). Although the strict Prisoner's Dilemma and Stag Hunt are fully contained in our original cube, it is clear that the volume corresponding to Deadlock, to Chicken, and to the Prisoner's Dilemma without the standard constraint continue beyond it.

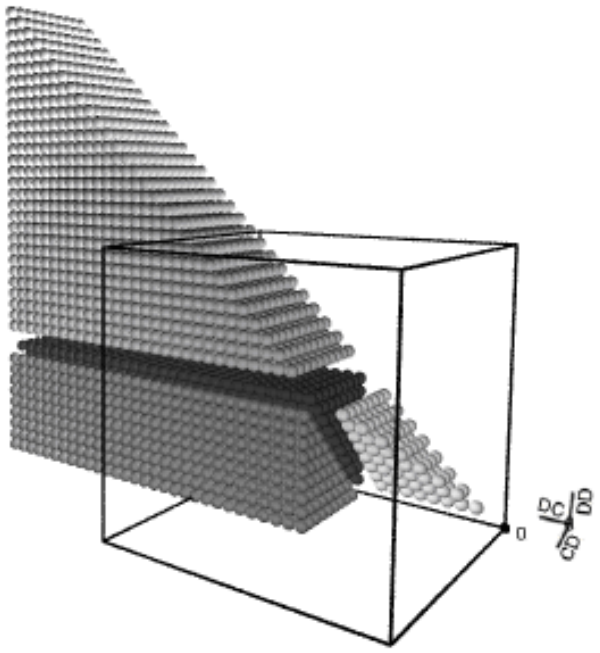

Figure 14. A view of the extended cube for DC and DD values greater than twice CC

IV. The Robustness of TFT

What we have proposed is a graphical map of the universe of 2 x 2 game theory. One thing such a map offers is a measure of robustness across changes in 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—that hold for a wider range of matrix values—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.[6]

With such a measure, we will also be able to offer a direct image of 'inclusive robustness'. A phenomenon X may hold at all matrix values at which Y holds, though phenomenon Y does not appear in all cases X does. Set-theoretic relationships of game-theoretic sub-phenomena, super-phenomena, union and intersection phenomena should be immediately obvious from their display in the cube. In this section and the next we offer two examples of the application of the matrix robustness measure.

TFT, we have noted, has a reputation as a robust effect across different forms of competition: Axelrod's round-robin tournaments (Axelrod 1980a, 1980b), Axelrod and Hamilton's replicator dynamics tournaments (Axelrod & Hamilton 1981), and in a spatialized competition of simple strategies (Grim 1995, 1996). 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. 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 for the Prisoner's Dilemma, it is well known that dominance first goes 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 against the background of All-D and C-then-All-D. Tit for Tat eventually conquers the entire array (Figure 15). A full evolution of this and later arrays can be seen at www.ptft.org/robustness.

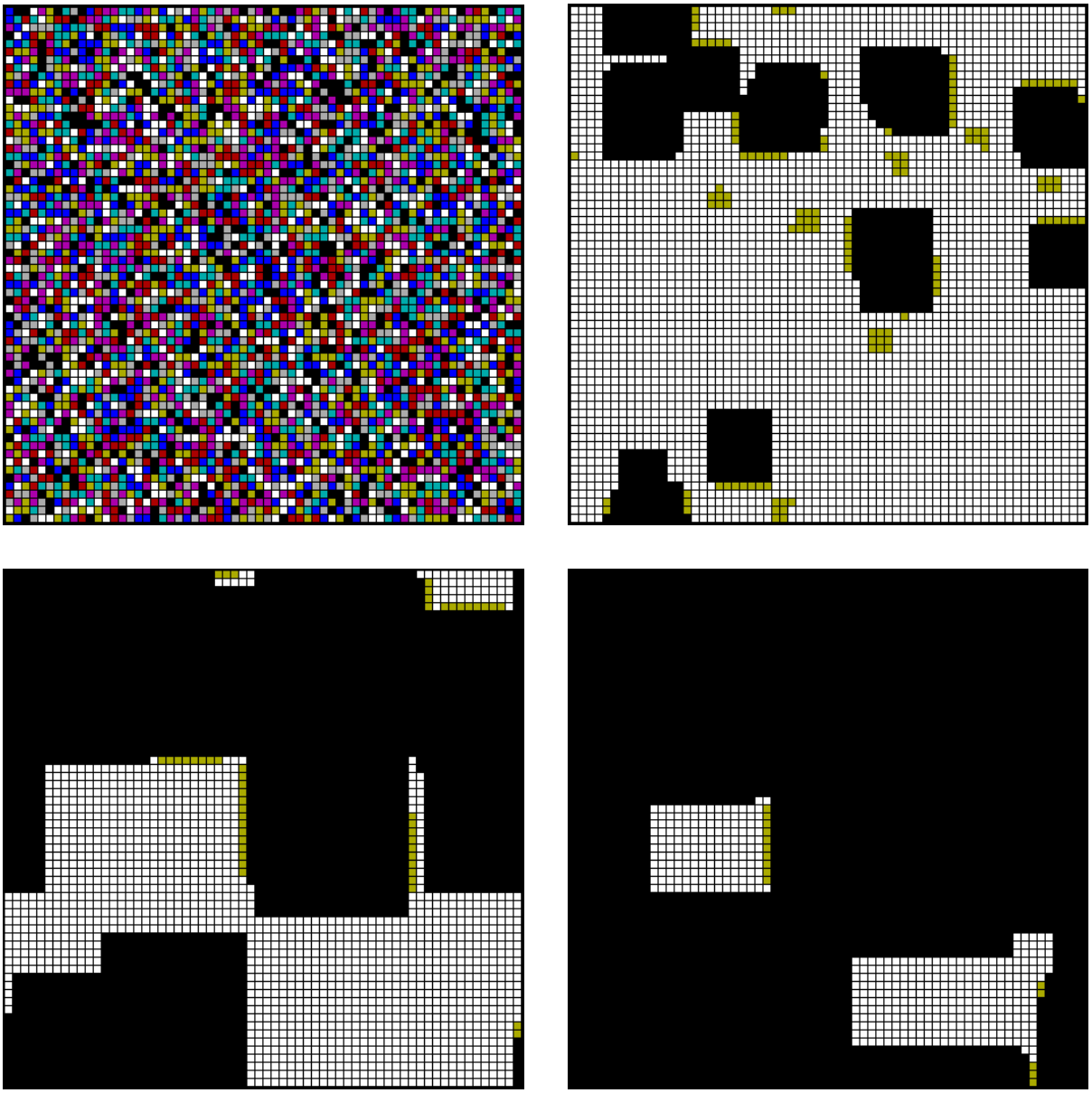

Figure 15. Conquest by TFT in a randomized environment of 8 reactive strategies

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

In order to answer that question, we took results across 8,000 spatialized competitions, using values for CC, 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.

When plotted, these points give us a clear indication of the robustness of the spatialized TFT effect across changes in matrix values (Figure 16). Results are shown from three chosen angles. A fully rotating image of the result can be found at www.ptft.org/paq/robustness.

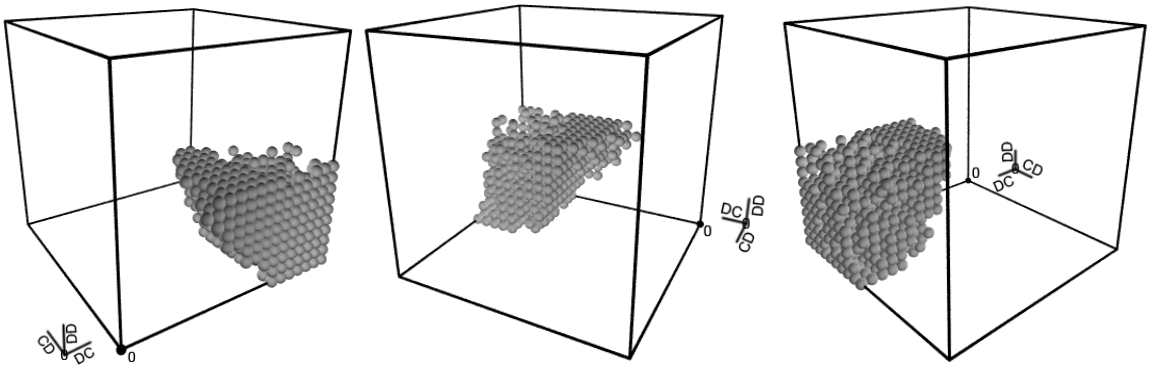

Figure 16. The spatialized TFT effect

Comparison with the extent of the Prisoner's Dilemma shown in Figures 2 and 3 indicates that the spatialized TFT effect appears through most of the area of the Prisoner's Dilemma, even when the extent of the game is broadened by dropping the CC > (CD + DC) / 2 condition. The TFT effect also appears beyond that area. In Figure 17, we graph only those matrix values for which the TFT effect appears that are *not* Prisoner's Dilemma values in even the broad sense. It can be seen that the effect spreads into a great proportion of Chicken (gray), a few values within Stag Hunt (light gray), and a cluster of values that fall under none of the standard games (shown in black). Conquest by TFT in a spatialized environment turns out to be an importantly robust effect across matrix values. In the next section we use this measure to compare the matrix robustness of this game-theoretic effect with another.

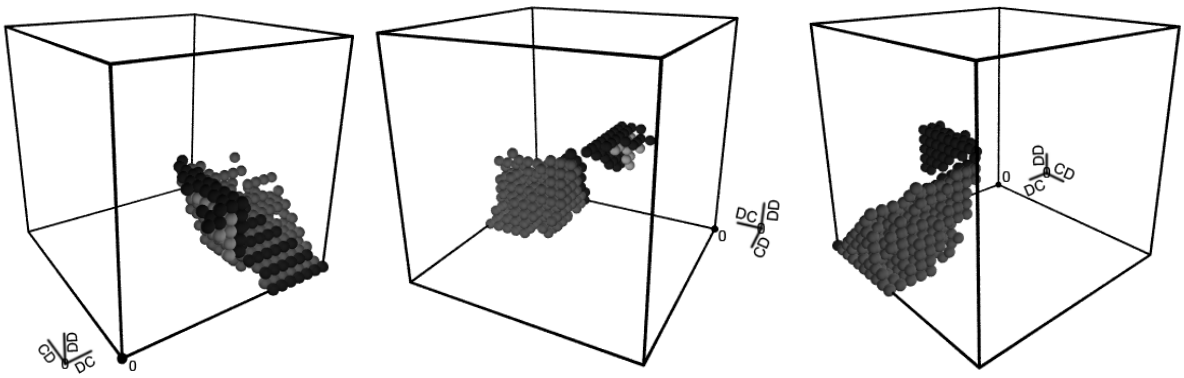

Figure 17. The extent of the spatialized TFT effect beyond the Prisoner's Dilemma

V. The Robustness of the Contact Hypothesis

In this section, we offer another effect for comparison: a game-theoretic instantiation of the contact hypothesis (as developed in Grim, Selinger, Braynen, Rosenberger, Au, Louie, & Connolly 2004; Grim, Selinger, Braynen, Rosenberger, Au, Louie, & Connolly 2005).

There are many theories regarding the nature and sources of prejudice in the social psychological literature, but only one major theory about how to reduce prejudice—the contact hypothesis. The contact hypothesis posits that under the right conditions, prejudice between groups will be reduced as those groups are integrated (Allport 1954; Pettigrew 1998; Zirkel & Cantor 2004). It has a range of empirical support and has played an important role in public policy starting with *Brown v. Board of Education*. A computational model for such a hypothesis, we suggest, would need to include at least the following features: (i) distinct groups, (ii) behaviors which may or may not be differentiated by actor and recipient groups, (iii) advantages and disadvantages resulting from these behaviors, (iv) an updating mechanism for behavior, and (v) configurations of greater and lesser contact between the different groups.

We have used game-theoretic resources to construct a model of this type: our model features cellular automata that play a spatialized version of the iterated Prisoner's Dilemma. Cells play only with their eight contiguous neighbors, and 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 18)—we are able to assess the success of PTFT in different environments. The contact hypothesis is tested by contrasting success the prejudicial strategy PTFT in the segregated environment with its success in the integrated one.

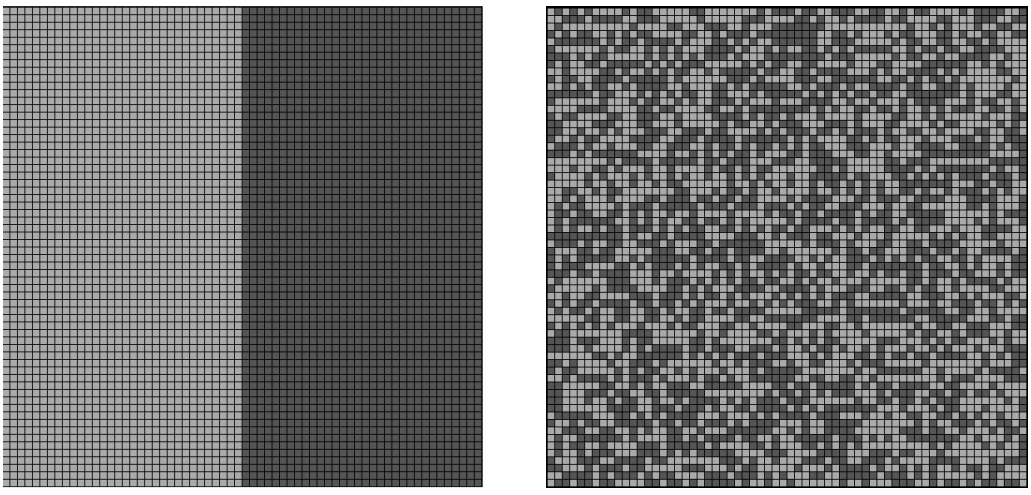

Figure 18. Segregated (left) and mixed patterns of background color

We find that 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 19, 20). In the mixed array, on the other hand, TFT eventually takes-over almost the entire space, leaving only very small clusters of the color-sensitive PTFT (Figures 21, 22). We claim that these results provide strong computational support for the contact hypothesis, and that social psychologists should pay closer attention to spatialized game-theoretic elements of advantage and disadvantage; these may indeed play a crucial role in the mechanism that facilitates prejudice reduction in contact situations.

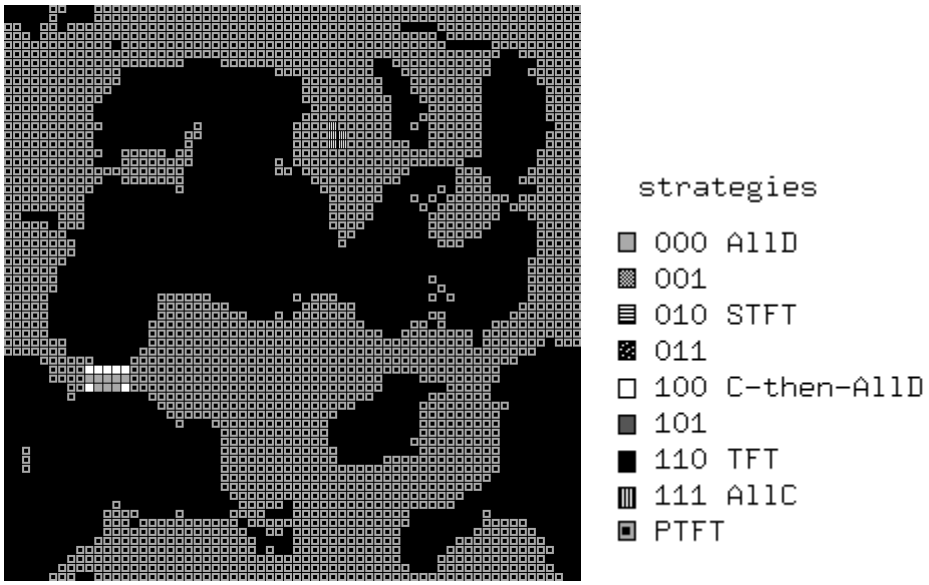

Figure 19. Evolution of randomized strategies to shared dominance by TFT and PTFT in an array segregated by color. A complete evolution can be seen at www.ptft.org/robustness.

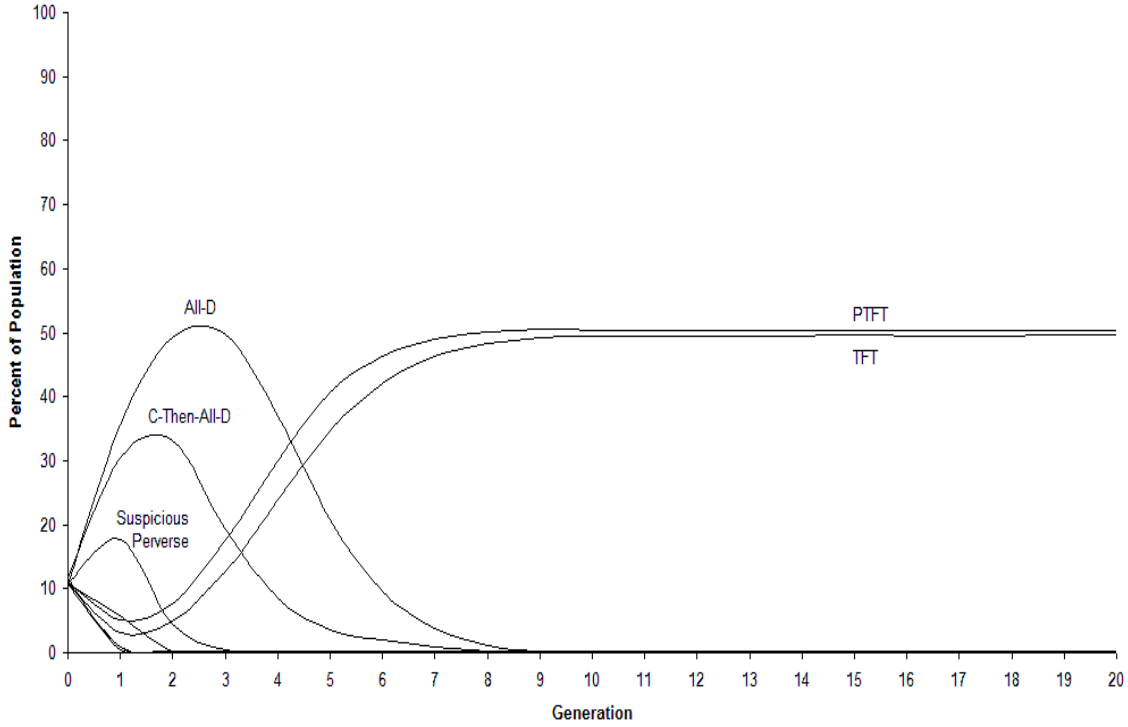

Figure 20. Percentages of the population for 9 strategies in an array segregated by color (20 generations shown)

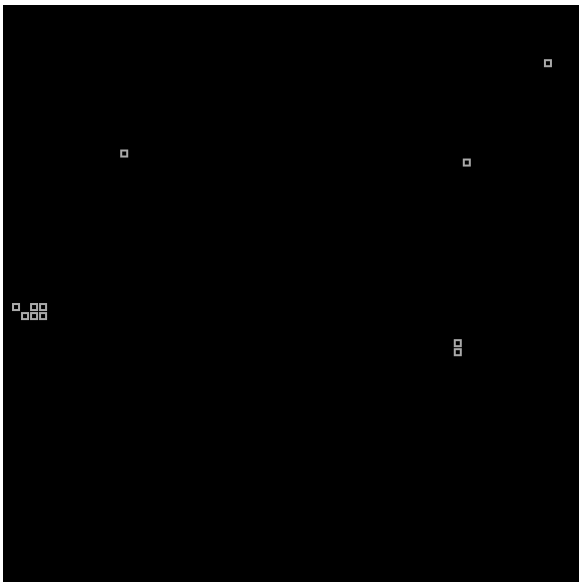

Figure 21. Evolution of randomized strategies to dominance by TFT in an integrated (randomized) color array

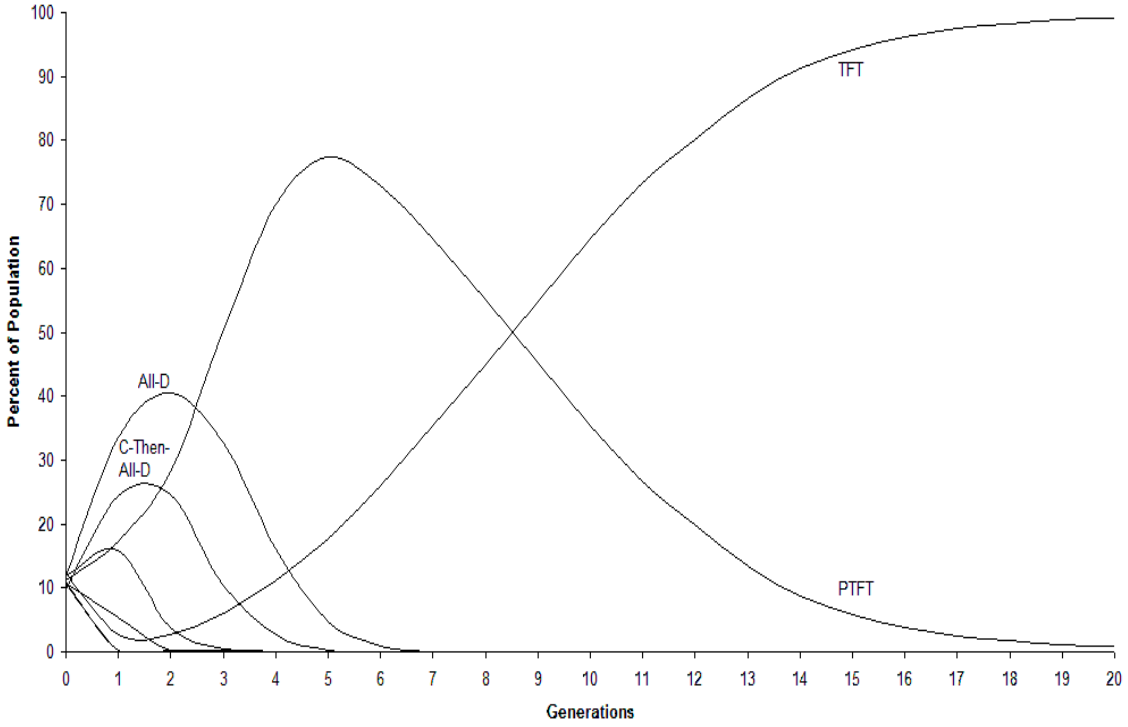

Figure 22. Percentages of the population for 9 strategies in an array randomized by color (20 generations shown)

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 contact 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 23 shows a graphic portrayal of the matrix robustness of the contact effect in these terms.

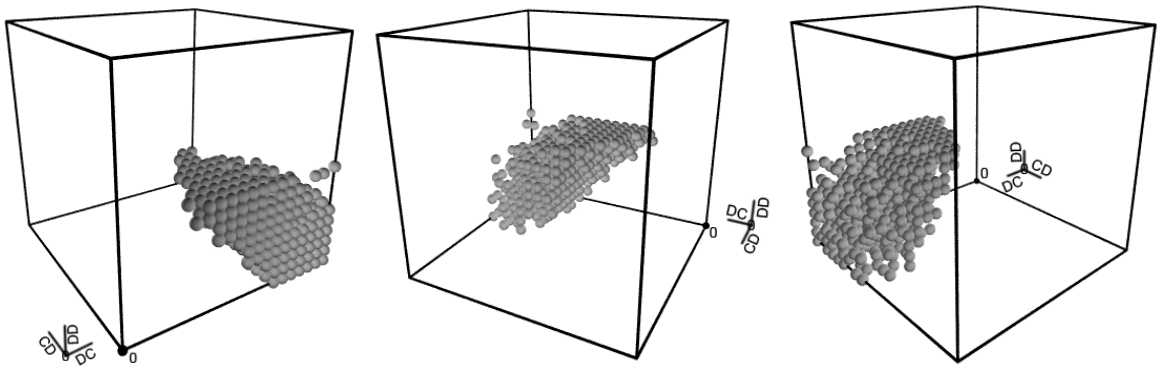

Figure 23. The PTFT effect

With two effects in hand, our measure allows a graphic comparison in terms of matrix robustness. Here as in the TFT effect, comparison with Figures 2 and 3 indicates that the PTFT effect is evident throughout the area of the Prisoner's Dilemma. It is, in fact, evident throughout the larger area of the Prisoner's Dilemma without the CC > [CD + DC] / 2 constraint. We can also compare the extent of the PTFT effect in Figure 24, however, with the extent of the spatialized TFT effect in Figure 16. That comparison vindicates the matrix robustness of the PTFT effect. TFT, we've 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.

The PTFT effect, like the TFT effect before it, extends beyond the limits of both the Prisoner's Dilemma proper and the larger area of the Prisoner's Dilemma without the standard constraint. In Figure 24, like in Figure 17, we eliminate that central area of the effect, showing the extent to which it similarly occupies a large area of Chicken (gray), a few matrices of Stag Hunt (light gray), and a cluster of values beyond any of the standard games (black).

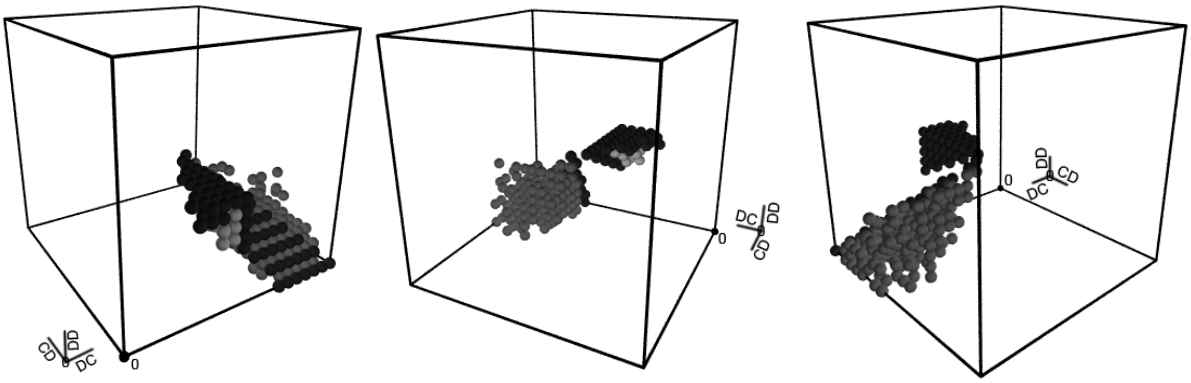

Figure 24. The extent of the PTFT effect beyond the Prisoner's Dilemma

VI. Conclusion

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

Some such measure, we think, is long overdue. Much of game theory has concentrated not only on the particular game of the Prisoner's Dilemma but on a specific set of matrix values for that game. Applications within theoretical biology, economics, and social and political philosophy quite often assume that the inequalities characteristic of the Prisoner's Dilemma can be taken as characteristic of biological, economic, or social life generally. It is common to move swiftly from that assumption to the specific values 5 > 3 > 1 > 0 with no argument at all. The widespread focus on this set of values has misled some into thinking that results established for that single matrix can automatically be taken as results regarding the Prisoner's Dilemma in general (for correctives see Nowak & May 1993; Lindgren & Nordahl 1994; Braynen 2004).

Relying upon any single set of matrix values can lead one to mistake a fragile and limited effect for a broad and robust one. It can also cause one to miss stronger effects in a wider neighborhood of values that do not happen to include one's chosen matrix point. A corrective for these dangers would be to accompany new results quite routinely with a measure of their robustness across matrix values.

We suggest that the game-theoretic cube instantiates effectively a measure of this sort. By selecting one normalized value, we provide the viewer with an opportunity to exploit what Herbert Simon called 'perceptual inferences' (Larkin & Simon 1987): matrix robustness can be envisaged in a rotating three dimensional display. In standardizing the measure across different models, the game-theoretic cube permits a direct comparison of the matrix robustness of different effects. This allows us to see at a glance that the areas of two effects are disjoint, intersect, or the area of one is a subset of the other.

There is also much to be learned from the comparison of different robustness measures. As outlined above, Axelrod and Hammond measure the robustness of their model by seeing whether the effects still occur when parameter values are doubled and halved (Axelrod & Hammond 2003). Gilbert and Troitzsch suggest another method: randomly sampling parameter values in order to chart variations in an effect. "Plotting the values of the outputs generated from many runs of the simulation will give an indication of the functional form of the relationship between the parameters and the outputs and will indicate whether small parameter changes give rise to large output variations" (Gilbert & Troitzsch 2002, 23). Axelrod and Hammond's technique has the advantage of economy: a relatively small number of additional runs are required. Unfortunately, theirs is also a fragile measure for robustness: since halving and doubling are relative to the initial values chosen, that initial choice may determine whether an effect is portrayed as robust or not. A similar problem will appear for any measure that relies on an algebraic variation on initial values. The Axelrod-Hammond test will also give false robustness positives, of course, for cases in which an effect holds at initial values, at half values, and at double values, but fails in the spaces in between. Gilbert and Troitzsch's technique avoids this latter difficulty, but becomes progressively less economical as a larger number of randomized values are tested. It remains fragile with regard to the ranges in which the randomized values are to be chosen.

The matrix robustness measure offered here, in contrast, is itself robust. Because it represents a sampling across all possible values, it avoids fragility in terms of either initial values or chosen ranges of random sampling. It must also be admitted, however, that it is a relatively expensive measure. As long as the effects at issue are those that appear in short runs, a survey across matrix values seems well worth the minimal cost. Where effects become complex, on the other hand, requiring runs that extend to weeks and months, a measure this comprehensive may become prohibitive.

It becomes clear in the very first steps of trying to analyze the concept that 'robustness' comes in many forms, senses, or types. The measure we have outlined is explicitly limited to robustness across variations in matrix values.[7] Even within spatialized game theory, it would be desirable to have a gauge of robustness across the structure of the spatialization, across updating mechanisms, and across changes in strategy sets as well.

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. In some cases it may be possible to apply the general strategy we have used here, with a measure that is itself robust because it represents or encapsulates all possible variations. In other cases it may not be possible. 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.

Notes

1. D'Arms, Batterman, and G∴rny 1998 discuss both representativeness and robustness as virtues of game-theoretic models, but do not note this important link between the two. They rightly note that there are importantly different kinds of robustness, and that "most authors who invoke robustness as an explanatory virtue are not very clear about what exactly makes the allegedly robust feature robust" (90).
2. This second condition, interestingly enough, legislates against an alternating strategy that shows up commonly in both experimental game theory and in everyday life, studied formally and with simulations in Vanderschraaf and Skyrms 2003.
3. "Game theorists often devote rather less attention to demonstrating that their games accurately model actual human interactions than one could wish…For better or worse, the prisoner's dilemma has been widely accepted among philosophers as teaching us something important about ordinary conduct" (D'Arms, Batterman, and G∴rny 1998, 89).
4. After developing the 3-dimensional model detailed here we discovered a 2-dimensional anticipation in Lindgren and Nordahl 1994 which also uses the trick of normalization. Their Fig. 10 is an attempt at an image across matrix values, though it captures only that part of the universe in which CC > CD and though their application is focused on the search for distinct cellular automata rules.
5. We are obliged to Paul St. Denis for calling 'shadows' to our attention and for insisting on their importance in the cube as a whole.
6. Relative measures of robustness in the cube, where one effect includes another, are fairly safe. Care should be taken in comparison of absolute volumes in different areas, however. In the full four-dimensional universe, the volume occupied by Stag Hunt and its shadow, for example, will be the same. In normalizing to a single CC value, as indicated above, this is not guaranteed to be the case. Normalization to values other than CC can also be expected to change the image of the game-theoretic universe considerably.
7. The Axelrod-Hammond and Gilbert-Troitzsch tests are somewhat more generalizable, because for example they call for halving and doubling or randomly sampling across various parameters. Many aspects of robustness, however—many dimensions of variation—extend beyond mere parameter values.

References

Allport, G. W.: (1954), The Nature of Prejudice, Addison-Wesley, Cambridge, Mass.
Axelrod, R.: (1980a), 'Effective Choice in the Prisoner's Dilemma', Journal of Conflict Resolution 24, 3-25.
Axelrod, R.: (1980b), 'More Effective Choice in the Prisoner's Dilemma', Journal of Conflict Resolution 24, 379-403.
Axelrod, R.: (1984), The Evolution of Cooperation, Basic Books, New York.
Axelrod, R. and Hamilton, W. D.: (1981), 'The Evolution of Cooperation', Science 211, 1390-1396.
Axelrod, R. and Hammond, R. A.: (2003), 'The Evolution of Ethnocentric Behavior', Midwest Political Science Convention, April 3-6, Chicago, IL.

Bergstrom, T.: (2002), 'Evolution of Social Behavior: Individual and Group Selection Models', Journal of Economic Perspectives 16, 231-238.

Brauchli, K., Killingback, T. & Doebeli, M: (1999), "Evolution of cooperation in spatially structured populations," Journal of Theoretical Biology 200, 405-417.

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

D'Arms, J., Batterman, R. and G∴rny, K.: (1998), 'Game Theoretic Explanations and the Evolution of Justice', Philosophy of Science 65, 76-102.

Gilbert, N. and Troitzsch, K. G.: (2002), Simulation for the Social Scientist, Open University Press, Buckingham.

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

Grim, P.: (1996), 'Spatialization and Greater Generosity in the Stochastic Prisoner's Dilemma', BioSystems 37, 3-17.

Grim, P.: (2005), 'Concrete Images for Abstract Questions: A Philosophical View', in Engström, T. and Selinger, E. (Eds.) Rethinking Theories and Practices of Imaging, forthcoming.

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 (2005), 95-125.

Harms, W.: (2001), 'Cooperative Boundary Populations: the Evolution of Cooperation on Mortality Risk Gradients," Journal of Theoretical Biology 213, 299-313.

Larkin, J. and Simon, H. A.: (1987), 'Why a Diagram is (Sometimes) Worth 10,000 Words', Cognitive Science 11, 65-99.

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

M ¯bius, A. F.: (1827) [1976], Der barycentrische Calcul, Leipzig (1827), Georg Ohms [1976], Hildesheim, Germany.

Nakamaru, M., Matsuda, H., and Iwasa, Y.: (1997), "The evolution of cooperation in a lattice-structured population," Journal of Theoretical Biology 184, 65-81.

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

Nowak, M., and Sigmund, K.: (1993), "Chaos and the evolution of cooperation," Proceedings of the National Academy of Sciences 90, 5091-5094.

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

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

Rucker, R.: (1984), The 4th Dimension: Toward a Geometry of Higher Reality, Houghton Mifflin, Boston.

Sigmund, K.: (1993), Games of Life, Oxford University Press, New York.

Skyrms, B.: (2001), 'The Stag Hunt', Presidential Address of the Pacific Division of the American Philosophical Association, Proceedings and Addresses of the APA 75: 31-41.
Skyrms, B.: (2004), The Stag Hunt and the Evolution of Social Structure, Cambridge University Press, New York.
Vandershcraaf, P. and Skyrms, B.: (2003), 'Learning to Take Turns', Erkenntnis 50, 311-348.
Wedekind, C., and Manfred, M.: (1996), 'Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus Generous Tit-for-Tat,' Proceedings of the National Academy of Sciences 93, 2686-2689.
Zirkel, S. and Cantor, N.: (2004), '50 Years After *Brown v. Board of Education*: The Promise and Challenge of Multicultural Education', Journal of Social Issues 60(1), 1-15.