Chapter 5

Game Theory's World in a Matrix

By the fall of 1969, the Americans and North Vietnamese had reached stalemate in their negotiations to end the Vietnam War. Richard Nixon had entered office the previous January promising to end the conflict, yet American GIs continued to return home in body bags. Meanwhile, at negotiations in Paris, the North Vietnamese steadfastly refused concessions to the Americans. Nixon exuded fury — and wanted to be sure the Vietnamese and their Soviet allies knew it. "I call it the Madman Theory, Bob," he told his chief of staff H.R. Haldeman, who would later serve time in the wake of the Watergate scandal. "I want the North Vietnamese to believe that I've reached the point that I might do anything to stop the war" — anything, including escalating to an all-out nuclear exchange.[1] Thus, toward the end of October, the air force launched the first waves of a massive airborne exercise, "Giant Lance," in which nuclear-armed B-32 bombers prowled the skies over the polar ice cap, towards and away from the Soviet Union. While their presence was designed to be apparent to Soviet military observers — who certainly would alert the Soviet leadership to the American threat — the flights were carefully hidden from the view of both the American public and allies around the world.[2]

Was this truly madness, or did it reflect a more calculated kind of lunacy? For nuclear strategists and political scientists of the 1960s, steeped in the rational-choice vernacular employed by the likes of Herman Kahn [Chapter 3] it must have seemed like

---

[1] H.R. Haldeman, with Joseph DiMona, *The Ends of Power* (New York: Times Books, 1978), p. 83. Quoted in Scott D. Sagan and Jeremi Suri, "The Madman Nuclear Alert: Secrecy, Signaling, and Safety in October 1969" *International Security* 27.4 (2003) 150-183.
[2] *Ibid*.

evidence of the latter. In particular, Nixon's bizarre logic of international relations has often been laid at the doorstep of game theory, a mathematical theory of interaction between "rational" individuals (defined in a sense peculiar to the theory) that had initially been developed by John von Neumann and Oskar Morgenstern in their 1944 book, *Theory of Games and Economic Behavior*. Kahn had made qualitative reference to the game of "chicken" in his analysis of escalation and deterrence, suggesting that calculated, rational choices could be made at each step of the escalation ladder, and during the 1960s formal elements of game theory became increasingly intertwined with discussions of war and peace. In his influential 1960 book, *The Strategy of Conflict*, the economist Thomas Schelling had drawn on game theory to suggest the employment of a "threat that leaves something to chance," the introduction of an element of randomness into the bargaining mix of carrots and sticks employed by a negotiator, since actually following through on a threat, or transparently failing to follow through, carries a strategic cost. The strategic use of randomness (the employment of so-called "mixed strategies" by game-players) was an essential ingredient of game theory from some of its earliest formulations by von Neumann in the 1920s.[3] As the "rational-actor perspective" pioneered by Schelling spread across academic political science in the following years, the Cold War thus was reinterpreted as a specific kind of game: a dilemma game in which rational calculation failed to achieve rational outcomes, and conversely, in which apparent "irrationality" was the only rational strategy. Subsequent histories of Cold War nuclear policy and

[3] John von Neumann, "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen* 100 (1928): 295-320.

intellectual culture, both popular and academic, reiterate the connection between game theory and this kind of thinking.[4]

The emergence and persistence of game theory at the center of debates over nuclear strategy, arms control, and international diplomacy presents a puzzle in light of the foregoing chapters. We have seen in Chapter Three how a rational choice approach to the study of nuclear strategy quickly proved empirically questionable during the 1950s and 1960s, and how a richer understanding of psychology, grounded in theories of "cognitive dissonance" was called for to supplement detached calculation as the basis for decision-making. Yet again and again, Cold War intellectuals turned to the spare notations and logic of game theory to tackle problems of strategy and arms control. In 1967, even as the limitations of rational calculation for solving such problems were becoming clear in the work of Charles Osgood and others, the Princeton consulting group Mathematica could present game theoretic models of the Vietnam War to their patrons at

ure 3.1: Nixon's dilemma in a matrix.

		V.C.	
		NOT ESCALATE	ESCALATE
U.S.	NOT ESCALATE	3 3 De-escalation	1 4 V.C. Victory
	ESCALATE	4 1 U.S. Victory	2 2 Escalation

[4] See especially William Poundstone, *Prisoner's Dilemma: John von Neumann, Game Theory, and the Puzzle of the Bomb* (New York: Doubleday, 1992); and Philip Mirowski, *Machine Dreams: Economics Becomes a Cyborg Science* (Cambridge: Cambridge University Press, 2002).

the Arms Control and Disarmament agency that would precisely capture the dilemma facing Nixon: whether to escalate the war in the quest for victory while possibly provoking Vietnamese escalation and an even bloodier stalemate. That the Cold War was literally a game in this stripped-down, spare sense remained a consistent point of departure for discussions of arms races and nuclear war, even as the adequacy of game theory's calculating brand of rationality came in for criticism.

To understand the persistence of such games in Cold War strategic debates (not to mention far-flung corners of the sciences) this chapter explores the several episodes in the history of the particular game captured in the matrix above, better known as the Prisoner's Dilemma (PD). PD is only one of 72 two-person non-zero-sum games: the game of "chicken" is another. Yet from its initial formulation by mathematicians working on behalf of the United States Air Force in the early 1950s, no other game has been so commonly associated with the paradoxes of security in the age of nuclear weapons. In the years since, PD has come to focus attention precisely on the divergence between the optimizing rationality embodied in mathematical programming models and the more complex kind of rationality needed to achieve substantively rational national-security outcomes. Despite the high hopes of its early practitioners, game theory *per se* did not provide anything like a "solution" to these problems or a promising calculus for thinking about "the problem of the bomb." If anything, from a very early date, PD suggested the failure of purely calculational approach to these issues.

More impressive than any practical results of game theory was the way that its notational devices and conceptual framework proved an exceptionally flexible and adaptable set of tools for coding and thinking about behavior in a wide range of

disciplinary contexts throughout the Cold War period. The game matrix in particular served as a key function in the research strategy of social and behavioral scientists in this period. As we saw in the previous chapter, postwar behavioral scientists found it profitable to focus on particular microcosms in which to study human behavior – different "situations," whether that meant a laboratory, a room glimpsed through a one-way mirror, an island, or a particular social encounter – but getting from observed behavior in the microcosm to general insights into the nature of human social interaction and choice behavior requires a certain narrowing of vision, a stripping away of aspects of a situation deemed non-essential, and the preservation of those that appear essential. This "x-ray strategy" allowed game-theoretic rationality to vault between contexts and across disciplinary lines and spatial and temporal scales. In the process, Nixon's Vietnam War became a PD game, played out between the superpowers — but so too did the interactions of human laboratory subjects, economic actors, and even insects undergoing natural selection. With these situations encompassed by a common game matrix, the peculiar amalgam of Cold War rationality could spread far beyond the contexts in which it was initially produced.

3.1. Game Theory and its Discontents at RAND

The earliest version of the game now known as the Prisoner's Dilemma was devised by mathematicians Merrill Flood (whose experiments on bargaining we have already encountered in Chapter 1) and Melvin Dresher at the RAND Corporation in January 1950. At that time they performed a brief experiment in which two players repeatedly played a game that they originally called "a non-cooperative pair," with

pennies for prizes.[5] By the spring of 1950, the game had acquired the story with which it is now commonly associated, given to it by the Princeton mathematician Albert Tucker who was trying to explain the game to an audience of psychologists at Stanford.[6] One of the earliest versions of this story — titled simply "A Two-Person Dilemma" was apparently written down by Tucker at Stanford in May of 1950. The story runs as follows:

> Two men, charged with a joint violation of law, are held separately by the police. Each is told that
>
> (1) If one confesses and the other does not, the former will be given a reward of one unit and the latter will be fined two units.
>
> (2) If both confess, each will be fined one unit.
>
> At the same time each has good reason to believe that
>
> (3) If neither confesses, both will go clear.[7]

Or, captured concisely in game theory's "payoff matrix" [**Figure 3.2**]:

I	II	Confess	not confess
confess		(-1, -1)	(1, -2)
not confess		(-2, 1)	(0, 0)

Here, the matrix entries (e.g., (-2, 1)) denote the payoffs (in some unit of value) to the row player and the column player, respectively. Thus the "Prisoner's Dilemma" came into existence in its familiar, symmetric form. The "dilemma" in question is often taken to refer to the decision facing the two prisoners: should I confess, locking in a 1-unit fine, or should I stay silent in hope of gaining freedom, only to risk a fine of two units if the other prisoner turns state's evidence?

Looking back to that winter of 1950 — the winter sandwiched between the first Soviet nuclear weapons test in August of 1949 and the outbreak of the Korean War in

[5] Flood, "Some Experimental Games" RAND RM-789-1 (20 June 1952), p. 17.
[6] Poundstone, *Prisoner's Dilemma*, p. 117.
[7] See Tucker to Flood, "A Two-Person Dilemma." Merrill Flood Papers, Box 1, folder "Notes, 1929-1967." Note that this is the story told on pp. 117-8 of Poundstone.

June 1950 — it is difficult not to see in PD the logic of the arms race and military escalation. Certainly, the image of a world divided, cordoned off as if into two separate prison cells — red team versus blue team, communists versus capitalists — resonates. So too does the challenge of uncoordinated action in a situation where both parties have immense potential to harm one another, with their only chance to cooperate and emerge unscathed from their cells dependent on a very fragile trust.[8] Yet the Prisoner's Dilemma did not immediately become the Cold War: rather, the dilemma that concerned the RAND mathematicians had less to do with the one facing the prisoners (or the superpowers), and more to do with the challenge this particular game posed for their attempts to build a theory of games that would be of service to their military patrons. Their forays into experimental behavioral science notwithstanding, Flood and Dresher were first and foremost mathematicians, formulators of axioms and provers of theorems. Moreover, the direction of their mathematical interests was intimately connected with the status of game theory as a branch of applied mathematics in the late 1940s, and the nature of the intellectual agreement reached between practitioners of game theory and the postwar Air Force.

Its appearance toward the end of World War II notwithstanding, game theory was initially developed largely outside of the military context over a period of nearly twenty years, culminating in von Neumann and Morgenstern's *Theory of Games and Economic Behavior*. In their book, Von Neumann and Morgenstern sought to establish games (like poker or chess) as the fundamental unit of analysis for a new social science. This new

[8] For analysis of the narratives surrounding the PD game, see Mary S. Morgan, "The Curious Case of the Prisoner's Dilemma: Model Situation? Exemplary Narrative?" in *Model Systems, Cases, Exemplary Experiments* eds. Angela N.H. Creager et al. (Durham: Duke University Press, 2007).

vision for social science (its creators hoped) would principally address what they saw as a number of shortcomings of traditional economic theory. Applying logic to mathematical axioms of "rational behavior" in game situations, von Neumann and Morgenstern's theory sought to "solve" games. In their view, a "solution" — "a characterization of 'rational behavior'" in a given game, would ideally consist of a "complete set of rules of behavior in all conceivable situations" (cf. Chapter 1 on the significance of "rules").[9]

Yet despite the hefty size of *Theory of Games* when it appeared in 1944, the only part of the theory that came close to realizing this vision dealt with two-person zero-sum games, that is, games in which the winnings of one player were the losses of the other. In such situations, the principle of "rationality" to be applied was fairly straightforward: choose a strategy that will maximize your expected winnings while simultaneously minimizing your opponent's expected winnings (a so-called "minimax strategy"). The key to "solving" such games was von Neumann's insight that a player's rational strategy might not follow a determinate course of action, but could choose a particular course of action at random according to a probability distribution. If these kinds of randomized strategies ("mixed strategies") were available to players, von Neumann could prove that "rational" strategies existed. If rationality meant maximization, the calculated use of randomness made rationality possible. Yet even in the relatively simple situation of the two-person zero-sum game, von Neumann only proved that solutions *existed*, rather than providing algorithms for the actual calculation of courses of action. The theory of games involving more than two players, of situations where bargaining over surpluses was possible, remained still more fragmentary. For these games, von Neumann and

[9] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1953), p. 33.

Morgenstern had suggested that the players would form coalitions to win and somehow divide the spoils of their collusion, subject to the constraint that individual players might "defect" to demand a greater share of the winnings elsewhere. This part of the theory provided nothing like the "complete set of rules of behavior" that von Neumann and Morgenstern had hoped for from the outset: solutions consisted of *sets* of possible payoff distributions to players; they provided little guidance to players on how to proceed; and von Neumann could not even prove that all games possessed such solutions (they don't).[10]

This situation was problematic because game theory's appeal to the military was built on the promise of solving games to determine rules of strategic interaction. As we saw in Chapter 2, this was also the promise held out by the mathematics of linear programming and numerical methods of optimization, which also attracted significant military funding and interest during this period. Game theory and linear programming are in fact closely related, as George Dantzig discovered when he visited von Neumann at Princeton in 1947. During their meeting, von Neumann conjectured that the problem of solving two-person zero-sum games and the linear programming problem were identical: in this instance, the game player, like the Air Force Office of Statistical Control, sought to maximize a linear function subject to a linear system of inequalities. This equivalence was further explored by Albert Tucker and his students at Princeton in subsequent years.[11]

---

[10] See e.g., W.F. Lucas, "The Proof that a Game may not have a Solution" RAND RM-5543-PR (January, 1968)

[11] On Dantzig's visit to Princeton and its aftermath, see e.g. Dantzig's introduction to E.D. Nering and A.W. Tucker, *Linear Programming and Related Problems* (Boston: Academic Press, 1993); George Dantzig, "Reminiscences about the Origins of Linear Programming," *Operations Research Letters* 1 (1982): 43-48; and Jan Lenstra, Alexander

The mathematics of the two-person zero-sum game thus became a key focus of attention for the community of mathematicians at Princeton and at RAND for a couple interrelated reasons: not only was problem of solving such games equivalent to practical problems of programming and logistics, but von Neumann had already developed a fairly coherent understanding of what it meant to "solve" such games for rules of rational behavior. Therefore the bulk of game theory studies pursued at RAND and elsewhere focused on finding solutions to particular two-person zero-sum games, such as models of duels between fighter and bomber aircraft, or games in which commanders had to allocate scarce resources across multiple battlefields, on the assumption that his opponent would make similar calculations.[12]

The connection between game theory and the military was further reinforced in the later 1940s by the development of computers and algorithms for actually finding practical solutions to two-person zero-sum games and linear programs. While we have seen that truly "optimal" solutions to such problems were not necessarily attainable given the computing state-of-the-art, the computer's calculational abilities remained a constant point of reference for those interested in solving games and related optimization problems. By 1950, this point of reference would be close at hand indeed: RAND had acquired a commercially available analog computer in 1948, and a year later, Corporation mathematicians began scouting the possibilities for constructing their own electronic

---

Rinnooy Kan and Alexander Schrijver, eds., *History of Mathematical Programming: A Collection of Personal Reminiscences* (Amsterdam: North-Holland, 1991).

[12] Cf. Philip Mirowski, "When Games Grow Deadly Serious: The Military Influence upon the Evolution of Game Theory" in *Economics and National Security: A History of their Interaction* (Durham: Duke University Press, 1991).

digital computer, which would become operational in 1953.[13] Computers could even be called on to mimic the human tactic of "bluffing" (or Nixon's unpredictable "madman" tactics) through the creation of algorithms for generating game theory's "mixed strategies," strategies where a game player chooses a course of action at random. Already in the spring of 1947, the RAND Corporation had developed a device that would convert inputs from a physical "random frequency pulse source" into randomly distributed digits printed on IBM punch cards. Within a few years, one of the mathematicians involved in the project could look forward to the day when improved numerical processes and improved computational power "will permit us to compute our random numbers as we need them." The computer – whether as concept or as material reality – had the potential to serve as game theory's ideal rational agent, bringing both mechanistic calculation and near-perfect randomness within a common ambit.[14]

Despite these successes there was a growing recognition that outside the computer-ready rationality of the two-person zero-sum game lay a great gulf in game-theoretical knowledge. As the RAND mathematicians noted in Research Memorandum after Research Memorandum throughout the late 1940s, von Neumann and Morgenstern's method for analyzing non-zero-sum games needed reassessment. One problem identified quite early concerned the formation of coalitions. As one mathematician put it already in

[13] On the history of computing hardware at RAND see Willis H. Ware, "RAND Contributions to the Development of Computing" at: http://www.rand.org/about/history/ware.html

[14] George W. Brown, "History of RAND's Random Digits — Summary" RAND P-113 (June 1949), p. 5. On random number production and its application to the solution of differential equations and the simulation of random processes at RAND and elsewhere see e.g., N. Metropolis, "The Beginning of the Monte Carlo Method" *Los Alamos Science* Special Issue (1987), pp. 125-129; Peter Galison, *Image and Logic: A Material Culture of Microphysics* (Chicago: University of Chicago Press, 1997), chapter 8; Sharon Ghamari-Tabrizi, *The Worlds of Herman Kahn: The Intuitive Science of Thermonuclear War* (Cambridge, MA: Harvard University Press, 2005), pp. 133-136.

a 1948 report on the state of game theory at RAND, "the assumption may be considered utopian" that players would form coalitions in many instances, and he called for the investigation of games in which there was no possibility of coalition formation.[15] But more troublesome still was the fact that von Neumann and Morgenstern's solution seemed incapable of prescribing "rational behavior" in the same way that the theory of the two-person zero-sum game had done so clearly. Their solutions, as Albert Tucker and Duncan Luce would write in 1959, "seem neither to prescribe rational behavior nor to predict behavior with sufficient precision to be of empirical value." The problem of practical reasoning – how to decide what one should do in any given situation – could not simply be replaced with rational calculation.[16]

As a result of these shortcomings, alternative "solution concepts" — alternate paths to the holy grail of "solving" a game — proliferated among the RAND-affiliated game theorists in the late 1940s and early 1950s. Perhaps the most sweeping attempt in this regard stemmed from the work of John Nash, then a Princeton graduate student who spent summers at RAND in the late 1940s and early 1950s. Nash's vision for game theory distinguished between theories of "cooperative games" (such as von Neumann and Morgenstern's) and "non-cooperative" games, in which players act "without collaboration or communication of any sort."[17] Instead of modeling the formation of coalitions (which would eventually break up anyway as each player clamored for his share of the gains from the collaboration) Nash assumed from the outset that individuals would apply to non-zero-sum games the same principle of rationality-as-optimization that

[15] Olaf Helmer, "Recent Developments in the Mathematical Theory of Games," (RAOP-16, 30 April 1948), pp. 16-18.
[16] *Contributions to the Theory of Games, Vol. IV,* eds. R. Duncan Luce and Albert Tucker (Princeton: Princeton University Press, 1959), p. 2.
[17] John F. Nash, "Non-Cooperative Games" (Ph.D. Dissertation, May 1950), p. 1.

had worked so well in the context of zero-sum games (where communication between the players was pointless). By this logic, players would seek a strategy that "maximizes [the player's] payoff if the strategies of the others are held fixed." The resulting set of strategies would constitute an "equilibrium point."[18]

Nash's solution concept — which he sent off for publication in the fall of 1949 — was a key piece of the context in which Flood and Dresher performed their first game experiments at RAND. Here is how Tucker continued the analysis of the PD in his memo from the spring of 1950:

> Clearly, for each man the pure strategy "to confess" dominates the pure strategy "not to confess." Hence, there is a unique equilibrium point given by the two pure strategies "to confess." In contrast with this non-cooperative solution one sees that both men would profit if they could form a coalition binding each other "not to confess." [19]

The term "equilibrium point" refers to Nash's non-cooperative solution concept, which would seem to suggest a strategy of mutual confession – thereby locking in a suboptimal outcome for both players. This stands at odds with the kind of solution von Neumann and Morgenstern might have proposed, i.e., "form a coalition binding each other "not to confess." (Or, as we saw in Chapter 1, to babysit.) Seen in this light, the principal "dilemma" of PD was not the one facing the prisoners, but the one facing the RAND mathematicians seeking to develop a comprehensive theory of multi-player and non-zero sum games. In effect, PD began its existence as a glorified mathematical counterexample.

It is not clear how successful the RAND experiment was in resolving this dilemma – that is, the dilemma of which solution concept to choose for solving non-zero-

[18] Nash, "Non-Cooperative Games," p. 3.
[19] See Tucker to Flood, "A Two-Person Dilemma." Merrill Flood Papers, Box 1, folder "Notes, 1929-1967." Underlining in the original.

sum games. Flood concluded that the experimental subjects showed "no tendency to seek as the final solution…the Nash equilibrium point," but neither did they cooperate in a straightforward manner.[20] The result seemed to please von Neumann, who had never felt the Nash equilibrium concept particularly interesting or appealing as a solution for game theory.[21] Nash, for his part, felt that the experiment did not constitute an adequate test of his equilibrium concept. His objections — recorded in Flood and Dresher's 1952 memo — hint at fundamental problems facing any attempt to experimentally verify any theory of games. To generate statistically meaningful data, the experimenters needed to repeat the game multiple times; however, since players have memories, subsequent games are effectively not the same game as earlier ones. One possible solution to this problem would be to have players rotate in and out of the game at random so that they could not get to know one another.[22]

But Nash's proposal begs the question: what was the point of a theory of games in the context of RAND and the needs of the Air Force? Was it intended to capture some essential feature of how people really do play games in some idealized and perfectly controlled situation that was probably impossible to create in a RAND Corporation office, much less on the battlefield? If not, then the point of further experimentation on PD would seem unclear. Perhaps not surprisingly, experiments on games appear to have tapered off by the mid-1950s in tandem with a decline in enthusiasm for game theory at the Corporation more generally. The decline doubtless had many causes, from budget cuts to the impact of the McCarthy security hearings on the RAND staff. However one

[20] Flood, "Some Experimental Games," p. 24.
[21] See e.g., Martin Shubik, "Game Theory at Princeton, 1949-1955: A Personal Reminiscence" in E. Roy Weintraub, ed., *Toward a History of Game Theory* (Durham: Duke University Press, 1992).
[22] Flood, "Some Experimental Games," p. 24.

cannot help but imagine that the methodological issues related to "solving" non-zero-sum games might have played a role as well. Game theory in the hands of its military patrons was intended as a guide to what *should* be: part of a program to improve (if not optimize) the use of weapons systems or the functioning of supply chains. Decision-making, accomplished by the solution of linear programs or tactical games via computer program or servomechanism, was intended to bypass the "human factor" as much as possible. Knowledge of how humans actually behave was not the object of interest, nor was it necessarily desirable.

3.2. Mathematics' Loss is Psychology's Gain

Mathematicians' project of "solving" games via optimization algorithms may have seemed headed toward a dead end in the 1950s, but meanwhile PD's disciplinary center of gravity began to shift from mathematics toward psychology, especially social psychology. At first blush, this shift seems odd given that psychology and game theory had had a somewhat antagonistic relationship dating to the founding works of game theory in the interwar period. At least in the case of the two person zero-sum game, Von Neumann argued that "it makes no difference which of the two players is the better psychologist" since the solution to the game could truly be reduced to calculation.[23] However, as we shall see here, the appeal of game theory to psychologists stemmed precisely from the fact that some of the precisely specified situations it analyzed — like PD — did *not* result in behavior that was predictable via any *a priori* criterion of optimization. PD thus offered a structured, controlled template for producing

[23] John von Neumann, "On the Theory of Games of Strategy" in *Contributions to the Theory of Games,* vol. 4, eds. A. W. Tucker and R.D. Luce (Princeton: Princeton University Press, 1959), p. 23.

psychological knowledge. Such knowledge — far more than the theories of the RAND mathematicians — would ultimately help insert PD into discussions of the dilemmas of international arms control in the thermonuclear age.

PD would form the basis for numerous studies of human behavior in the 1950s and 1960s [**Figure 3.3:** Room layout for experimental games, 1959], most notably Anatol Rapoport's exhaustive experimental exploration of the game, *Prisoner's Dilemma: A Study in Conflict and Cooperation*, published in 1965. Rapoport was in many ways the ideal person to bridge the gap between game theory as practiced at RAND and psychology during this period: a mathematics Ph.D., he also spent time at the University of Chicago's Committee on Mathematical Biology in the late 1940s and early 1950s, eventually moving to the University of Michigan in 1955. Rapoport first encountered game theory during a year's sabbatical in 1954-55 at the Center for Advanced Study in the Behavioral Sciences at Stanford, where he encountered PD in a seminar led by the mathematician and measurement theorist R. Duncan Luce. According to his autobiography, Rapoport immediately saw the implications of the game for thinking about patterns of conflict and cooperation, both among individuals and nations. Upon his arrival at Michigan the next year, he embarked on a multi-year study of human teamwork and cooperation for the Air Force, which was interested in improving the performance of its flight crews. In the course of these studies, Rapoport began to use experimental subjects' behavior in PD to quantify their tendency to cooperate as team members.[24]

Rapoport's 1965 book offers a fascinating analysis of the relevance of game theory to psychology. The theory of the two-person zero-sum game contained little of

[24] See Anatol Rapoport, *Certainties and Doubts: A Philosophy of Life* (New York: Black Rose Books, 2000), chapters 8-9.

interest to the psychologist, he argued, since the rational course of action in such situations appeared straightforward. Such games were only potentially interesting to the extent that actual human behavior might "irrationally" depart from the theoretical predictions.[25] However, "[b]esides the irrational deviations, the psychologically interesting aspects of conflict are those stemming from mixed motives, where the confrontation is not only between the conflicting parties but also between the conflicting motives within each participant." In this regard the study of non-zero sum games was significant not as a path to developing a successful "theory" of such games; rather, "the potentially rich contributions of game theory to psychology will derive from the failure of game theory rather than from its successes."[26] The Prisoner's Dilemma game illustrated precisely this "failure" of game theory to develop a coherent understanding of rationality, since "the paradox remains unresolved as long as we insist on adhering to the concept of rationality which makes perfect sense in zero-sum games but which makes questionable sense in nonzero-sum games."[27] Here, Rapoport was clearly referring to Nash's equilibrium solution for nonzero-sum games, which suggested a strategy of mutual non-cooperation in PD that Rapoport indeed felt was of questionable rationality. Failure of this kind of theory was thus a necessary prerequisite for developing insights into "real psychology": "the realm of personality, intellect, and moral commitment." Psychology thus embraced all aspects of human reasoning, where game theory focused on a much narrower, calculating kind of rationality.[28]

---

[25] Anatol Rapoport and Albert M. Chammah, *Prisoner's Dilemma; A Study in Conflict and Cooperation* (Ann Arbor: University of Michigan Press, 1965), p. 6.
[26] Rapoport and Chammah, *Prisoner's Dilemma*, p. 11.
[27] Rapoport and Chammah, *Prisoner's Dilemma*, p. 13.
[28] Rapoport and Chammah, *Prisoner's Dilemma*, p. vi.

Rapoport's findings, however, actually shed very little light on these issues — at least when it came to individuals. Like Flood and Dresher before him, Rapoport focused principally on analyzing multi-play runs of PD by the same players (rather than, for example, comparing the behavior different populations of players in a single play of the game). In this setup, despite a complete ban on communication between the players in all the trials, the effects of repeated interactions and the payoffs at stake seemed more significant than preexisting characteristics of the individual players (such as intelligence or personality) in determining outcomes. Rapoport did analyze the differences between the ways males and females played the game; yet even here the *pairing* of the players (MM, FF, MF) seemed more relevant than anything else to understanding the patterns of cooperation and conflict that emerged. "Whatever individual differences exist among the players (and it is difficult to believe other than that they exist) tend to be ironed out in the course of the interactions between them," so that much of the variation in outcomes "is accounted for not by the inherent propensities of the players to cooperate or not cooperate, but rather by the characteristic instabilities of the dynamic process which governs the interactions in Prisoner's Dilemma."[29] While there was a slightly greater tendency overall toward cooperation than non-cooperation (with the overall frequency depending mostly on the structure of the payoffs involved), the most impressive result of the trials was an overwhelming tendency toward *conformity*: players either cooperated most of the time or did not cooperate, so that "Typically, toward the end of the sessions over ninety percent of the responses are matched."[30]

[29] Rapoport and Chammah, *Prisoner's Dilemma*, p. 198-9.
[30] Rapoport and Chammah, *Prisoner's Dilemma*, p. 199.

If Rapoport focused much of his attention on interaction processes rather than personal qualities, the opposite was true of another pioneer of PD laboratory studies, Morton Deutsch. Deutsch's much-cited experimental study of PD, "Trust and Suspicion," appeared in *The Journal of Conflict Resolution* in 1958, making it one of the first experimental PD game studies to appear since Flood and Dresher's 1952 paper. As with Rapoport, Deutsch represented a very different disciplinary lineage than Flood, Dresher, and Tucker, completing his Ph.D. thesis in experimental social psychology at MIT's Research Center for Group Dynamics in 1948.[31] From its roots in the work of Kurt Lewin and his students in the 1930s, Group Dynamics had focused on developing an experimental understanding of the interaction between individual personality and social environment, most notably the relationship between personalities and leadership styles on the one hand, and group productivity in the workplace and in civic life. Lewin's classic study during this period presented observations of social interaction in two groups of fifth- and sixth-grade children who were brought together around craft activities. Comparing observations on the behavior of the groups under different styles of adult leadership — one "democratic," participatory, and consultative, and the other "authoritarian," characterized by top-down leadership — Lewin clearly thought he had found convincing evidence for the superiority of democratic leadership. The authoritarian group exhibited greater social tension, hostility, and scapegoating behaviors; the democratic group was not only characterized by greater intra-group communication and stability, but also compiled a better work record.[32]

---

[31] Erica Frydenberg, *Morton Deutsch: A Life and Legacy of Mediation and Conflict Resolution* (Brisbane: Australian Academic Press, 2005), p. 56.
[32] Kurt Lewin and Ronald Lippitt, "An Experimental Approach to the Study of Autocracy and Democracy: A Preliminary Note" *Sociometry* 1.3/4 (January-April 1938), pp. 292-

During the 1940s and 1950s this tradition of research attracted substantial support from both the military and industry, which valued insights into teamwork, and by reformers interested in resolving social conflicts. Correspondingly, the goal of research in Group Dynamics was the use of motivational training techniques to induce social and behavioral change, whether on the factory floor, in combat teams, or in housing developments.[33] During the 1950s, Deutsch in many ways epitomized this intellectual tradition. In addition to holding a faculty position, during 1952-54 he was a member of the Committee on Civil Rights of the Society for the Psychological Study of Social Issues, in connection with his work studying interracial housing in New York and Newark, New Jersey.[34] He also received funding from the Office of Naval Research (ONR) for his experimental research into conditions promoting cooperation in small groups.[35]

Deutsch's landmark 1958 study of trust and suspicion clearly drew on his work for the ONR and focused on understanding the conditions that would foster trusting attitudes in a small group setting. "Trust" in this instance was not simply a matter of cognition, of successful prediction of future events, but also involved the positive and

---

300; see also Kurt Lewin, Ronald Lippitt, and Ralph K. White, "Patterns of Aggressive Behavior in Experimentally Created 'Social Climates'" *The Journal of Social Psychology, S.P.S.S.I. Bulletin* 10 (1939), pp. 271-299.

[33] On Lewin, see Marvin Weisbord, *Productive Workplaces Revisited* (2004) chapters 4-5; Alfred J. Marrow, *The Practical Theorist: The Life and World of Kurt Lewin* (New York: Basic Books, 1969, 1984); William Graebner, *The Engineering of Consent: Democracy and Authority in Twentieth-Century America* (Madison: University of Wisconsin Press, 1987).

[34] Frydenberg, *Morton Deutsch*, p. 58. See also Morton Deutsch and Mary Evans Collins, *Interracial Housing: A Psychological Evaluation of a Social Experiment* (Minneapolis: The University of Minnesota Press, 1951).

[35] See M. Deutsch, *Conditions Affecting Cooperation* (Final Technical Report for the Office of Naval Research, Contract NONR-285[10], February 1957), cited in Morton Deutsch, "Trust and Suspicion" *The Journal of Conflict Resolution* 2.4 (December 1958), pp. 265-279.

negative "motivational consequences" of confirmation or disconfirmation of belief. Deutsch hypothesized several factors that might increase the "individual's confidence that his trust will be fulfilled" relating to the perception the individual had of others. These included "the nature of the intentions that the individual perceives his potential object of trust to have; the perceived power of the object of trust to cause the desired events; the power relationship between the individual and his object of trust; the influence of communication upon the development of trust; the influence of third parties upon the development of trust between two people; the individual's self-esteem as it affects his readiness to trust."[36] Experimental plays of the Prisoner's Dilemma game, he argued, would provide the perfect opportunity to test the conditions that might reinforce trusting behavior since "The essential psychological feature of the game is that there is no possibility for 'rational' individual behavior in it unless the conditions for mutual trust exist."[37]

Here, Deutsch quite naturally equated "cooperation" and trust with rational behavior, with outcomes that are best for the "team" of players as a whole, and "motivational consequences" with the psychological impact on individuals of wins and losses. With these equivalences in place, Deutsch thus proceeded to test several possible factors creating trust. For example, he performed experimental trials of the game under three different "motivational orientations," "cooperative," "individualistic," and "competitive"; each orientation was conveyed before play via verbal instructions to the subjects "which characterized…the objectives they were to have in playing the game and

[36] Deutsch, "Trust and Suspicion," p. 269.
[37] Deutsch, "Trust and Suspicion," p. 270.

the objectives they could assume their co-player would have."[38] Perhaps not surprisingly, the "co-operative" orientation instructions produced consistently high percentages of cooperative strategy choice, whereas a "competitive" orientation was nearly always lowest.[39] Other experiments and observations focused on behavior in situations where communication was permitted or not permitted, with Deutsch observing that even players who were given the opportunity to communicate often did not do so effectively, in Deutsch's opinion. Ultimately he posed a question for future research: "How can communication opportunities be used to raise the individual's confidence that his trust will be fulfilled and also used to elicit trustworthy or responsible behavior?"[40]

Deutsch's analysis thus is striking in the richness of social interaction and social roles that he sought to investigate, drawing in considerations of cooperation, communication, power, and social connectedness. This sense of rationality, far richer than the rationality-as-optimization pursued by the RAND mathematicians or indeed the conformity discovered by Rapoport, is nevertheless in some ways oriented toward similar ends. Even if the goal was not to axiomatize reason, to reduce it to a set of rules and calculations, experts were still needed to engineer the motivational environment in which groups of individuals could come to behave rationally (in this case, to cooperate). Rationality would be generated not by computer but by some kind of collective therapy. But in the process of adapting games to the laboratory and to practical problems of mediating social conflicts, Deutsch's work became almost completely divorced from the "theory" of games in any sense that Flood and Dresher might have recognized.

[38] Deutsch, "Trust and Suspicion," p. 270.
[39] Deutsch, "Trust and Suspicion," p. 272.
[40] Deutsch "Trust and Suspicion," p. 273.

Rapoport and Deutsch both began their psychological investigations of PD with funding from the U.S. military, which was interested in understanding phenomena of teamwork and cooperation in small groups like the crews of airplanes or submarines. Quite quickly, however, both men came to imagine that the insights generated by their research had relevance to the arms race shaping up between the United States and the Soviet Union in the 1950s. Deutsch had been involved with antiwar causes since learning of the nuclear bombings at Hiroshima and Nagasaki. As a result, his doctoral thesis on learning in cooperative and competitive environments (which would set the scene for much of his work on trust and suspicion), had its roots in the postwar years when he "had been more interested in world peace than in education," his thesis's ostensible subject.[41] Rapoport's revelation was due in no small part to his personal convictions: a socialist, he had been a vocal opponent of the increasingly violent exchanges of rhetoric between the US and the Soviets since the late 1940s. As a result, during the 1954-55 year, he was also part of a reading group that that met to discuss the work of Lewis F. Richardson, a Quaker meteorologist who had brought to bear statistical and mathematical models to study the progression of arms races and the outbreak of wars. Among other things, Richardson had written down differential equations describing the dynamic interactions between nations undergoing arms buildups. Depending on the parameters of the equation, increased arms expenditures in one country could lead to increased arms expenditures in the other, with overall armaments crescendoing in a chain of reactions and counter-reactions.[42] Richardson's models took a page from the equations of classical physics, so that war, driven by moods for which "there are no rational components" is

[41] Frydenberg, *Morton Deutsch*, p. 67.
[42] See Anatol Rapoport, "Lewis F. Richardson's Mathematical Theory of War" *Conflict Resolution* 1.3 (Sept 1957), pp. 249-299.

like a disease with "regular, almost predestined course."[43] Nevertheless, according to Rapoport, "The connection between [PD] and the situation produced by the arms race occurred to me at once"; "cooperation" meant undertaking arms control, while non-cooperation meant continued weapons development.[44]

The PD game thus seemed a promising tool for investigating problems of conflict and cooperation at the international and interpersonal levels, a far broader mandate than Flood and Dresher's at RAND. This connection would be reinforced by the development of new institutions and sources of funding in the later 1950s that sought to apply the results of behavioral science to understanding Cold War problems of peace and violent conflict. These included the University of Michigan's Center for Research on Conflict Resolution (and the in-house *Journal of Conflict Resolution*, in which both Rapoport and Deutsch published extensively) and after 1961, the Arms Control and Disarmament Agency, which funded research in game theory throughout the 1960s.

In this context, PD *did* become a key theoretical framework for thinking about "the problem of the bomb," and the game matrix could leap from the mathematics of optimization to problems of war and peace writ large. However, the lesson of the theory would ultimately prove ambiguous for both Deutsch and Rapoport, as it did for many other academics associated with peace research and conflict resolution. What role would their knowledge of games play in resolving conflict? Game theory's spare description of reality allowed them to move seamlessly between human subjects in a laboratory and the affairs of nations — or so they thought. Clearly it did not promise a clear-cut procedural rationality capable of addressing the problems of social conflict and the arms race.

[43] Rapoport, "Lewis F. Richardson's Mathematical Theory of War," pp. 284-285.
[44] Rapoport, *Certainties and Doubts,* p. 113.

Deutsch's work sought to empower the group psychotherapist or counselor to act upon attitudes of the parties to a conflict and somehow guide them toward rational (i.e. cooperative) behavior. Rapoport, by contrast, would prove less enthusiastic about such attempts to "engineer" rationality, preferring to focus on using PD to demonstrate the possibility for (but not the necessity of) an enlightened, more empathetic logic of individual decision.[45] But in the end, neither set of insights seemed realistically capable of guaranteeing a cooperative outcome to the Cold War arms race, even if they seemed promising on the laboratory level. The arms race seemed to call for more durable solutions.

3.3. Rational Outcomes without Intelligent Actors

By the mid-1960s, once again, a new site for deliberating the problems of reason and violent conflict emerged in the study of animal behavior. The problem of the arms race would be solved not simply by exploring the recesses of our minds; more likely, its roots lay far back in our evolutionary history, or perhaps even in the logic of life itself. The shift from a cultural to an innately biological explanation of violence and aggression in humans and animals alike was perhaps surprising given the dominant state of scientific opinion in the 1950s. Among biologists interested in animal behavior, a long intellectual tradition emphasized the natural origins of social cooperation rather than competition. Darwin's "struggle for existence" did not operate only or even primarily on the level of individuals; altruistic behavior, wherein individuals sacrificed themselves for the good of others, also evolved through its benefit to the species as a while. The most fundamental

[45] See e.g., Anatol Rapoport, *Strategy and Conscience* (New York: Harper and Row, 1964).

drives in nature were toward social harmony and cooperation.[46] Aggression and competition certainly did occur in nature — for example, in fights for territory and mates — but they were always carefully restrained to prevent needless slaughter. Indeed, ethologists would argue that in many instances, animal fights had become so "ritualized" as to simply provide symbolic displays of threats, rarely resulting in serious violence.[47] Moreover, responding to postwar revelations of the horrors of Nazi racial science and eugenics, a number of prominent biologists came to question the connection between genetics and social behavior in humans, reversing the earlier emphasis of eugenics that had sought to address social problems through monitoring and manipulation of human heredity.[48]

Several developments changed this situation in the 1960s. In the realm of popular culture, a slate of works by authors such as Robert Ardrey and Desmond Morris popularized new theories of the origins of human social behavior, including violence. For example, drawing on earlier work by the anthropologist Raymond Dart, Ardrey's *African Genesis* (1961) argued that in fact a growing lust for hunting, bloodshed, and weaponry had driven the evolutionary transition from apes to humans.[49] In addition, a

---

[46] See e.g., Paul Crook, *Darwinism, War, and History: the Debate over the Biology of War from the "Origin of species" to the First World War* (Cambridge: Cambridge University Press, 1994); Gregg Mitman, *The State of Nature: Ecology, Community, and American Social Thought, 1900-1950* (Chicago: University of Chicago Press, 1992).
[47] See e.g., the review paper by Sir Julian Huxley, "Introduction: A Discussion of Ritualization of Behavior in Animals and Man" *Philosophical Transactions of the Royal Society of London* Series B (Biological Sciences) 251.772 (29 December 1966), pp. 249-271.
[48] On this shift within Eugenics see e.g., Daniel J. Kevles, *In the Name of Eugenics: Genetics and the Uses of Human Heredity* (New York: Knopf, 1985).
[49] Robert Ardrey, *African Genesis; A Personal Investigation into the Animal Origins and Nature of Man* (New York: Atheneum, 1961); see also Robert Ardrey, *The Territorial Imperative; A Personal Inquiry into the Animal Origins of Property and Nations* (New York: Atheneum, 1966); and Desmond Morris, *The Naked Ape; A Zoologist's Study of the Human Animal* (New York: Dell, 1967).

new generation of evolutionary theorists — most notably W.D. Hamilton, John Maynard Smith, George Price, Robert Trivers, and Richard Dawkins — emerged who were committed to restoring what they saw as Darwin's original emphasis on inheritance coupled with individual advantage as the engine of evolutionary change. This intellectual movement was reinforced by a new vision of life for the DNA age: organisms are information-processing machines, programmed by instructions coded in their genes from conception.[50] These intellectual movements raised challenges both for those who would privilege social interactions over biology as a force shaping behavior, and for those who explained adaptations by reference to innate social tendencies or to their contribution to the "survival of the species." The problem of how to reconcile the neo-Darwinian emphasis on individual advantage with the behavior of collectives thus lay at the heart of much evolutionary theorizing in the 1960s and 1970s. The Prisoner's Dilemma game (and game theory more generally) emerged in biology at precisely this time in association with debates over the evolutionary origins of altruism and aggression.

Before it could be deployed in the context of evolutionary biology, game theory had to be substantially reworked to apply to non-human actors. Humans play games for money and pleasure. By contrast, neo-Darwinian evolutionary theorists needed to find new metrics for the analysis of the costs and benefits arising from evolutionary adaptations — essentially, a "utility function" for life itself. The value of evolutionary adaptations would not be measured by their benefit to species or even to individual organisms; they were preserved if they helped to perpetuate the genes that controlled them. This perspective was pioneered by the British biologist William D. Hamilton in

[50] See e.g., Lily E. Kay, *Who Wrote the Book of Life? A History of the Genetic Code* (Stanford: Stanford University Press, 2000); Evelyn Fox Keller, *The Century of the Gene* (Cambridge, MA: Harvard University Press, 2000).

several papers from 1963 and 1964 that created the theory of "kin selection." Here, Hamilton directly tackled the problem of how to explain altruistic behavior in terms of its evolutionary advantage to individuals. Organisms did not dispense altruism out of concern with collective solidarity; rather, such behavior evolved if its benefit to an individual organism's inclusive fitness — a quantity that included the survival benefit to other organisms weighted by their degree of genetic relatedness — outweighed the costs to the individual.[51]

Second, theorists like Hamilton refocused attention away from the struggle between organisms and their environment and toward competition within a population, between individuals of a species. It was in this context that Hamilton first introduced ideas from game theory into his work in a 1967 paper that sought to explain the evolution of the sex ratio, especially the existence of lopsided sex ratios in a number species of insects. Sex ratios did not emerge to maximize the reproductive success of the species as a whole; rather, to a greater or lesser degree of realism, organisms behaved as if they played games with each other in which the sex ratio in their offspring represented their "strategy" in the game. Since natural selection favored individuals with greater fitness relative to the rest of the population, organisms were locked into a zero-sum game with one another. Therefore they would evolve as if they selected two-person zero-sum game theory's "minimax strategy" — or as Hamilton put it, an "unbeatable" strategy against which no other player could do better. Depending on the structure of the population in

[51] W.D. Hamilton, "The Genetical Evolution of Social Behavior, I," *The Journal of Theoretical Biology* 7 (1964), pp. 1-26; "The Genetical Evolution of Social Behavior, II," *The Journal of Theoretical Biology* 7 (1964), pp. 27-52.

question (e.g. the degree of reproductive mixing that its ecology permitted) species would develop different sex ratios.[52]

Organisms thus had the interests and competitive drive to be game players. But while Hamilton could use the language of "games" "strategy" and "choice" in talking about the behavior of tiny wasps and mites in 1967, he was clearly uncomfortable about the implications of this language. Some organisms certainly seemed to be capable of sensing their environments and adjusting their behavior accordingly. At the same time, imputing intentionality or the conscious ability to respond to environmental stimuli seemed pretty close to attributing to species precisely the kind of "collective interests" that neo-Darwinian theory sought to banish. Indeed, as Hamilton would remark in a letter to his friend and colleague George Price, while "[w]ith the tiny animals discussed I think it extremely unlikely that they are able to play the suggested 'game' intelligently, or recognize their own 'sex-ratio types'" nevertheless "I think there is an interesting theoretical problem as to how sex ratio behaviour should be expected to evolve if an intelligent animal like man was to find itself in the situation described."[53]

In this regard, Hamilton was particularly intrigued by Rapoport's analysis of the Prisoner's Dilemma game, which seemed to highlight the divergence between the behavior of animals with "exceptional intelligence" like humans, and those without. By 1967, Rapoport could suggest that collective and individual rationality in PD could be reconciled if one considered the problem within a broader framework of "metalogic," where players could consider strategies conditional upon the different possibilities for

[52] W.D. Hamilton, "Extraordinary Sex Ratios," *Science,* New Series, 156.3774 (28 April 1967), pp. 477-488.
[53] Hamilton to Price, 21 March 1968 (Item KPX1_5.5 Price Papers, Hamilton Archive, British Library)

how other players in the game might act.[54] However, Hamilton suggested, "against this the model in my paper seems to show that the proposed 'solution' does not hold for the animals discussed under natural selection, and I am doubtful whether intelligence makes much difference to the kind of solution that is possible. I am sure that prisoner's-dilemma situations are common and important in biological evolution."[55] More broadly, Hamilton spent much of his intellectual energy in the late 1960s musing on whether the evolution of reasoning abilities (in the form of speech, memory, and cognition) would do much to resolve a PD-type paradox. Perhaps the ability to communicate would also bring with it the ability to lie? And perhaps deceit would only be enhanced in a cognitive arms-race, as organisms evolved ever subtler and more complex ways to deceive one another? In the end, he would conclude that perhaps culture and the altruistic "values of civilized man" simply formed a "higher hypocrisy" intended to fool our fellow human beings.[56] There seemed to be no way to bridge the selfishness of genes with the kind of social cooperation envisioned by Rapoport and others approaching cooperation from a psychological perspective.

The perfect opportunity for Hamilton to share his musings on this problem arose in May of 1969 when he was invited to participate in a major interdisciplinary conference, held at the Smithsonian Institution in Washington D.C., that sought to explore the lessons of recent developments in ethology and the study of animal behavior for understanding the problem of human violence on both the national and international levels. Perhaps not surprisingly, Hamilton's contribution to the conference provided little

[54] Anatol Rapoport, "Escape from Paradox" *Scientific American* 217.1 (July 1967), pp. 50-56.
[55] Hamilton to Price, 21 March 1968 (Item KPX1_5.5, Price Papers, Hamilton Archive, British Library)
[56] See e.g., Hamilton to Price, 21 March 1968, Price Archive, Item KPX1_5.5.

solace for those who hoped to find a biological basis for altruism on either side of the human-animal divide. The basis for Hamilton's pessimism was a model of the possible pairwise interactions between individuals in which two genetic strategies are possible: one associated with a "normal" gene, and the other with a mutant "selfish gene" (M), with gene frequencies *p* and *q* respectively. Any given type of interaction could be represented by a 2×2 game matrix, with "payoffs" denominated in abstract units of evolutionary fitness [**Figure 3.4:** Hamilton's payoff matrix]. However, instead of trying to solve this game for the probabilities with which the players would adopt the two strategies (as a game theorist interested in human behavior would have done), Hamilton instead computed a difference equation that described how these "probabilities" — in this situation, the gene frequencies *p* and *q* — would evolve. If the game payoffs are those of a PD game, and even if the "selfish gene" is initially rare, Hamilton's difference equation suggested that it would spread through the population over time. Hamilton saw in this result a fundamental lesson of evolutionary biology: it is not as important to choose an "optimal" strategy as it is to choose a strategy that is simply better than those employed by other organisms in the same population. From an evolutionary perspective, it does not pay to cooperate in PD situations.[57]

While the result seemed decisive and coincided so clearly with Hamilton's intellectual instincts, its complete contradiction of Rapoport's results nevertheless bothered him, and the balance of the paper reads as a long deliberation on the complicated and unsettled relationship between game theory as it appeared in psychology on the one hand and Hamilton's vision for evolutionary theory other the other. Hamilton

[57] Hamilton, "Selection of Selfish and Altruistic Behavior in some Extreme Models," in Eisenberg and Dillon, eds., *Man and Beast: Comparative Social Behavior* (Washington, DC: Smithsonian Institution, 1971), pp. 59-91.

recognized that the theory of games between *human* players "presupposes being able to think, and, potentially, to communicate." This cognitive-psychological aspect of game theory made the theory a convincing model of human conflict and cooperation, but it also led to the recognition of irreconcilable conflicts between individual and group interests — interests that he felt had no clear counterparts in nonhuman populations, since animals did not possess the same subtleties of communication and cognition found in humans. Perhaps games needed to be solved in different ways for humans and non-humans, with something approximating Rapoport's "cooperative" solution for humans and Hamilton's "selfish" solution for non-humans. In such a case, Hamilton's solution would have little relevance to the human problem of "how it is rational to act" when thrust into a prisoner's dilemma, despite the fact that natural selection "has made us almost all that we are."[58]

Two encounters would change Hamilton's outlook over the next decade. The first was a chance meeting at the Washington conference between Hamilton and Robert Trivers, then a graduate student working with Harvard primatologist and anthropologist Irven DeVore. At DeVore's suggestion, Trivers had begun investigating the relevance of anthropological literature on "reciprocation" — the exchange of favors and aid — to explain the emergence of altruistic behavior among primates and among non-human organisms more generally. Hamilton's paper at the Washington conference came as a revelation to Trivers. In his presentation at the conference, Hamilton added a coda to his written paper in which he gestured to existing work on repeated PD games and suggested that a repeated-game framework might demonstrate the feasibility of cooperation as a solution to Hamilton's PD game. Trivers immediately saw the relevance of such games

[58] Hamilton, "Selection of Selfish and Altruistic Behavior," pp. 82-83.

to his problem of reciprocated altruism.[59] Shortly thereafter, in a groundbreaking paper titled "The Evolution of Reciprocal Altruism," he would note that "The relationship between two individuals repeatedly exposed to symmetrical reciprocal situations is exactly analogous to what game theorists call the Prisoner's Dilemma….Iterated games played between the same two individuals permit each player to respond to the behavior of the other." The extent of the communication going on in Trivers' models was questionable given that none of his mathematical formulations required individual organisms to remember their interactions with others or modify their behavior in the present in response to their expectations about the future. Indeed, this was critical to Trivers' argument since, following Hamilton's theory of kin selection, his goal was to "take the altruism out of altruism," that is, to remove the necessity for altruists to have intentions or choice. However, drawing on calculations suggested by Hamilton, he was able to show that in populations of organisms where players would interact a given number of times, it was possible for altruistic genes to maintain themselves against invasion from selfish genes, even if they might not spread from a low initial probability. Trivers proceeded to identify numerous examples of reciprocal altruism: mutualistic grooming, altruistic alarm calls in birds that alert the flock to the approach of predators, and so forth.[60]

The second encounter that would change Hamilton's mind about the possibility for cooperation evolving in repeated PD games came when he left Imperial College for the University of Michigan in 1978. By this point, both Anatol Rapoport and the Center for Research on Conflict Resolution had long since moved to other institutions.

---

[59] Hamilton to Wilton S. Dillon, Smithsonian Institution, 30 January 1970 (Hamilton Papers, British Library; no number yet assigned).
[60] *Quarterly Review of Biology* 46 (1971) pp. 35-57.

However, shortly after his arrival in Ann Arbor, Hamilton became acquainted with the political scientist Robert Axelrod, who was then implementing a series of computerized "tournaments" between strategies for repeated two-player PD submitted by game theorists across the country. The reigning champion strategy — submitted by none other than Rapoport himself — was "tit for tat," in which one player would reward the other's cooperation in one move with cooperation in the next (and likewise, punish non-cooperation with non-cooperation) [**Figure 3.5:** TIT FOR TAT against different populations].[61] Axelrod's style of investigation clearly appealed to Hamilton, especially since the computerization of strategies could banish unrealistic assumptions about the kind of interactions the players could have, reducing appeals to the memory, cognition, or reasoning ability of the fully mechanized game "players." The ultimate upshot of Hamilton and Axelrod's Ann Arbor meeting was an award-winning 1981 paper, "The Evolution of Cooperation," which demonstrated that tit-for-tat solutions could be found to PD if the chance of the players meeting again was high enough. The article would later provide the title for Axelrod's 1984 collected essays on cooperation in PD games, *The Evolution of Cooperation*.[62]

It is interesting to note that Axelrod begins *The Evolution of Cooperation* by invoking Hobbes' famous description of life in the state of nature that existed prior to the establishment of governments: "solitary, poor, nasty, brutish, and short." Clearly Axelrod saw repeated games and reciprocation as a starting point for explaining the

[61] On the PD tournaments see Axelrod, "Effective Choice in the Prisoner's Dilemma" *Journal of Conflict Resolution* 24 (1980), 3-25; "More Effective Choice in the Prisoner's Dilemma" *Journal of Conflict Resolution* 24 (1980), 379-403.
[62] Robert Axelrod and William D. Hamilton, "The Evolution of Cooperation," *Science* 211 (1981), pp. 1390-1396; and Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).

emergence of human institutions, social norms, and morality — concepts the evolutionary biologists had little use for. Moreover, Axelrod had no problem with granting human game-players the ability to pursue goal-directed, deliberate action, the neo-Darwinians' bête noire. But his starting point was nevertheless the state of nature and the randomly generated groups of blind, unthinking gene-machines that Hamilton and others insisted upon as the basis for their evolutionary models. Even as reciprocation and repeated PD games would again receive significant attention within political science in the 1980s as a result of Axelrod's work, the legacy of game theory's sojourn in evolutionary biology would remain. Cooperation in PD would emerge not from conscious reasoning or rational calculation, but from the evolutionary dynamics of selfish genes.

3.4. Conclusion: Writing Matrices, Locating Rationality

From mathematical optimization problems to the study of insect populations to reflections on the Cold War arms race itself, game theory's characteristic matrix notation helped intellectuals of this period to find Prisoner's Dilemmas almost everywhere. Examinations of reasoning, rationality, and choice were no longer totally bound to the specifics of particular situations, but could jump from one setting to another in tandem with transfers of research programs, investigatory techniques, careers, and insights. And as a result, a set of themes that characterized the amalgam of Cold War rationality could spread from specific contexts across a much broader intellectual landscape. Most notably, the rule-following computer appears throughout this history as a point of reference for exploring rational conduct, even as it simultaneously suggested the fundamental unreasonableness of Cold War rationality. The perfect Cold War rational

agent – whether embodied in John Nash's ideal experimental game-playing subjects or in Axelrod and Hamilton's simulated game-playing organisms – may have possessed prodigious calculational abilities, but he also lacked a set of faculties and qualities classically associated with reasoning: memory, choice, consciousness, goal-directedness, and even intelligence. This loss was alternately derided (for instance, by the social psychologists) or embraced (by the evolutionary biologists) in particular contexts and for particular reasons, yet it remained a central feature of rationality until at least the 1980s.

The apparent "solving" of repeated PD games via strategies of reciprocation preceded the end of the Cold War by only a few years. Just how connected the two developments were is open to debate. Certainly Axelrod's work gained a wide audience in political science and the arms control community; "Reciprocation" and "reciprocity" dovetailed with buzzwords common even among top-level policymakers since the advent of stepwise arms limitation talks in the 1970s. Yet in many ways, the gap between substantively rational outcomes and the procedures that would ensure them remained as vexing as they had been in the days of Flood and Dresher's experiments. Cooperation might be possible, but rational calculation would not necessarily get you there. Indeed, if it did emerge, a substantively rational outcome of PD might have less to do with the intelligence or decision-making abilities of the players and more with the blind historical processes that guided them. As with Adam Smith's "invisible hand," which ensured productive economic behavior without requiring good intentions on the part of the butcher or the baker, a cooperative conclusion to the Cold War need not have emerged from actors who were reasonable in any meaningful sense.

And, as cognitive psychologists increasingly came to assert in the 1970s and 1980s, this was probably for the best. As it turns out, humans in this era would prove to fall far short even of the narrow standards presumed by Cold War rationality. This growing recognition, emerging in tandem with the breakup of the Soviet Union and the dissolution of the national security consensus in the United States, would ultimately bring about the splintering of Cold War rationality and the privileging of "irrationality" as the prime characteristic of human decision-making.