

**ECONOMIC RESEARCH REPORTS**

***MODELING FREE CHOICE  
IN GAMES***

by

**Steven J. Brams**

**RR # 95-11**

**May 1995**

**C. V. STARR CENTER  
FOR APPLIED ECONOMICS**



**NEW YORK UNIVERSITY  
FACULTY OF ARTS AND SCIENCE  
DEPARTMENT OF ECONOMICS  
WASHINGTON SQUARE  
NEW YORK, N.Y. 10003**

## *Abstract*

### **Modeling Free Choice in Games**

The free and independent choices that game theory assumes are, on occasion, constrained by the game form postulated and its associated rules of play. Thus, if a game in normal (or strategic) form has a unique Nash equilibrium in pure strategies, the standard theory presumes rational players will choose it. But changing the rules of play to allow players to think ahead and move and countermove from from such an equilibrium may lead to a different equilibrium outcome, according to the “theory of moves” (TOM).

It turns out that of the 78 distinct  $2 \times 2$  strict ordinal games, there are exactly six, subsumed by a generic “Freedom Game,” in which the more far-sighted rationality calculations of TOM induce a player with a dominant strategy to depart from the unique pure-strategy Nash equilibrium associated with it and switch to its dominated strategy, which in turn induces the other player also to switch its strategy. The outcome that results is a “nonmyopic equilibrium” (NME)—based on a set of rules that makes the order of play endogenous and assumes players will not cycle—that benefits one and sometimes both players.

A second set of new rules, which assumes that cycling is possible, affords the players still more freedom of choice. However, this freedom may be limited by the exercise of “moving power,” which is the ability of one player to outlast the other player in the move-countermove process. Both sets of rules not only help one understand real-life choices players have actually made—some of which proved surprising—but also facilitate modeling dynamic processes that, by expanding the possible moves of players, enhance their freedom of choice.

# Modeling Free Choice in Games<sup>1</sup>

## 1. Introduction

A tenet of noncooperative game theory is that players are free to make any choices they like, as long as they are permitted by the postulated game form and its associated rules of play. In an extensive-form game defined by a game tree, for example, a player can choose any branch at a node where it is his or her turn to move next. In a normal (or strategic)-form game defined by a payoff matrix—wherein strategies describe plans of action that may be contingent on the choices of other players—a player can choose any strategy.

This freedom of choice is presumed channeled, however, by the rational calculations of the players. That is, while free to make any choices they please, players are, nevertheless, assumed to make moves, or choose strategies, that lead to better rather than worse outcomes. A major question in game theory is how to define “better” and “worse.”

I sidestep this issue by analyzing games in which better and worse are, on the surface, uncontroversial. These are two-person normal-form games of complete information in which one player has a dominant strategy. The other player, anticipating the first player’s choice of this strategy, has a best response, which leads to a Nash equilibrium.

It turns out that in some of these games, the player with the dominant strategy can do better by choosing his or her dominated strategy, provided he or she can do so before the other player has the opportunity to move or choose a strategy. Despite the dependence of the outcome on the order of play in such games, most game-theoretic models do not make the order of play *endogenous*—that is, the subject of rational calculation by the players.<sup>2</sup> Rather, the rules generally prescribe that one player moves first, or the players move simultaneously.

---

<sup>1</sup>I thank the C. V. Starr Center for Applied Economics at New York University for its support.

<sup>2</sup>There are some recent exceptions, including Hamilton and Slutsky (1990, 1993), Rosenthal (1991), and van Damme and Hurkens (1993). Typically, these models allow a player in the preplay phase of a game to choose when it will move in the play of the game. However, the choice of when to move applies only to a player’s initial strategy choice, whereas the nonmyopic calculations to be developed here assume that players, starting at outcomes, make moves and countermoves that depend on thinking several steps ahead.

Chicken is perhaps the best-known example of a game in which the order of play matters. If the players make simultaneous choices, as is assumed in the usual  $2 \times 2$  representation of this game, they may suffer their mutually worst outcome if neither “chickens out” by not swerving. But if one player (say, a male) can convince the other player (a female) that he will not swerve, he will obtain his best payoff, given that she responds optimally by swerving. Her optimal response, however, garners her only her next-worst payoff, so there is a disadvantage to a player’s moving second, rather than first, in Chicken.<sup>3</sup>

In this paper, I begin by assuming that players make simultaneous choices in a  $2 \times 2$  “generic game,” which I call the Freedom Game.<sup>4</sup> This game is *generic* in the sense that it subsumes several  $2 \times 2$  strict ordinal games, in which each of the two players has two strategies and is able strictly to rank the resulting four outcomes from best to worst (i.e., there are no ties).

The Freedom Game is defined by two conditions:

- (1) one player, whom I assume is Row (R), has a dominant strategy, which guarantees that the game has a unique Nash equilibrium in pure strategies (Hamilton and Slutsky, 1993, p. 50, Lemma 1);<sup>5</sup>
- (2) R has an incentive to depart from his dominant strategy and choose

---

<sup>3</sup>For reviews of the literature on the effects of different rules or protocols—including the timing of moves—on game outcomes, see Hirshleifer (1985) and Brams (1994).

<sup>4</sup>Elsewhere I call this game the “Surprise Game” (Brams, 1995)—for reasons I will spell out later—but here I give this game a very different interpretation and, hence, a new name. The nomenclature of “Freedom Game” is intended to be ironic, because the standard rules of play in fact restrict the choices that rational players would make in this game, compared with those they would make under two sets of new rules (to be described). While implications of the first set of rules were analyzed in Brams (1995), their interpretation here is new; both the implications and the interpretation of the second set of rules have not been previously analyzed for the Freedom Game.

<sup>5</sup>If neither player has a dominant strategy, there are either multiple Nash equilibria or no Nash equilibria in pure strategies. For example, Chicken has two pure-strategy Nash equilibria; it also has a mixed-strategy equilibrium if the payoffs are not ordinal (as assumed here) but cardinal utilities. In games like Chicken that have multiple Nash equilibria, there is an equilibrium-selection problem, which is the subject of the so-called refinements literature in game theory that seeks to narrow down (“refine”) the set of Nash equilibria to some proper subset, preferably with only one element (van Damme, 1991). By contrast, here I seek not to delimit Nash equilibria, since there is only one in the games studied, but instead to determine when a unique dominant-strategy Nash equilibrium will *not* be chosen.

his dominated strategy, given that he has the opportunity to move first in a 2 x 4 game.<sup>6</sup>

On first blush, this seems surprising, because R's dominated strategy is unconditionally worse than his dominant strategy.<sup>7</sup>

Of the 78 distinct 2 x 2 strict ordinal games (Rapoport and Guyer, 1966), R has an incentive to move first in six games, which break down into two classes. In three of the six games (class I), C also does better when R moves first, so it is in her interest to allow this to happen.

In the other three games (class II), there is a conflict of interest between the two players, because C does better moving first herself, or at least maintaining the status quo of simultaneous play in the 2 x 2 game. Consequently, if who moves first is left endogenous by the rules of play, then the structure of payoffs in class II games leaves uncertain what will happen.

I have incorporated this ambiguity into the theory of moves, or TOM (Brams, 1993, 1994; Brams and Mattli, 1993), which adds a dynamic component to classical game theory. Instead of assuming that players choose strategies—either simultaneously or sequentially—in normal-form games, TOM assumes that play begins at outcomes in the matrix, which are called *states* because they may be only temporary in nature. Players then decide whether to move or not move from these states by making rational calculations (to be described later) in order to reach “nonmyopic equilibria” (NMEs).

One theoretical finding of this paper is that the six 2 x 2 games in which R has an incentive to move first are the only games of the 78 in which a player would, according to TOM, move from a unique dominant-strategy (i.e., for one player) Nash equilibrium. It is in this sense that these equilibria are *unstable*: each player would, thinking ahead about its rational moves

---

<sup>6</sup>Henceforth, I assume R is male and the column player (C) is female.

<sup>7</sup>To be sure, R's dominated strategy in the 2 x 2 game is not dominated in the 2 x 4 game that results when R moves first. In fact, it is R's best response to C's (now-dominant) strategy in the 2 x 4 game, as I will show in section 3. But this rationale for R's switching from his dominant to his dominated strategy in a 2 x 2 game is cumbersome, as compared with making the order of moves endogenous in an explicitly dynamic theory.

and countermoves, move from the Nash equilibrium to an NME that is a different state.<sup>8</sup>

The one-to-one correspondence between games in which the Nash equilibria and NMEs differ, on the one hand, and dominant-strategy games in which who moves first matters, on the other, illustrates how TOM captures order-of-play effects directly in its concept of an NME. This relieves one of the task of having to ask whether, when the order of play is endogenous, the dominant-strategy Nash equilibrium in the 2 x 2 game is unstable. It is if and only if the players would move from this equilibrium to a different NME.

But the fragility of Nash equilibria in the six 2 x 2 games is not just a function of the nonmyopic calculations that lead to NMEs. Under a radically different set of rules of play, these Nash equilibria are no better predictors of where the players might end up.

The key difference between the first set of rules and the second set is that the second set allows the players to cycle. In two of the six games, the outcome that occurs under the second set depends on which, if either, player has “moving power,” which roughly means having the ability to force the other player to stop when that player has the next move. In the other four games, the outcome does not depend on which player has moving power.

Remarkably, whether moving power is effective (first two games) or irrelevant (second four games), the outcome induced is never the Nash equilibrium; it may or may not be the NME (depending on who has moving power). Thus, once again, the Nash equilibrium is suspect as a predictor of behavior when players are given free rein to make choices in a game.

## 2. Freedom *Not* to Be Surprised or Forced to Concede

The unexpected tie-in of unstable Nash equilibria on the one hand, and alternative rules that predict different outcomes on the other, is more than just a theoretical issue. Not only have these different outcomes actually occurred, but when they did they were the occasion of great surprise. As a case in point, the Japanese surprise attack on Pearl Harbor in 1941, which

---

<sup>8</sup>Technically, a Nash equilibrium is not a state but the strategies of players that produce this state. For convenience, however, I will usually identify equilibria—both Nash and nonmyopic—by their state (or payoffs at this state) rather than by the strategy pair that yields this state.

Brams (1995) modeled as a class II 2 x 2 game in which the Japan had an incentive to seize the initiative by attacking—at the expense of the United States—even though its dominant strategy in the simultaneous-play game was not to attack, shocked the world.

Taking the initiative may involve peaceful as well as belligerent actions, as illustrated by the second case analyzed in Brams (1995), Egyptian president Anwar Sadat's 1977 offer to visit Jerusalem to negotiate a peace treaty directly with the Israelis. The situation Sadat and Israeli prime minister Menachem Begin faced was modeled as a class I 2 x 2 class in which, unlike the Pearl Harbor game, both players benefited when the player with the dominant strategy (i.e., Sadat) moved first. That abandoning this strategy was risky for Sadat, but *not* irrational, seems borne out by his initial success yet later assassination.

Both the Japanese attack and Sadat's offer were astonishing events at the time. Such surprises, in my opinion, are well modeled in terms of a player's (perfectly rational) abandonment of a dominant strategy in favor of a dominated strategy, which in turn induces a countermove on the part of the surprised player to a different state, or NME.<sup>9</sup>

In laying bare the rationality of these moves in these cases, TOM helps to make them explicable and, therefore, unsurprising, at least in retrospect. Prospectively, TOM might help decision makers avoid "anticipation failures," caused by incomplete information about what rules are applicable. By making endogenous who moves first, TOM introduces a flexibility into thinking about how different scenarios might arise; by assuming that players think ahead, it provides a calculus for tracing out the dynamic moves they are likely to make over time, including when play will terminate.

On occasion, however, players may be less able to determine an endpoint than be convinced that they can outlast the other player. Then they might well choose to cycle. Indeed, in two of the six 2 x 2 games, this is a rational choice: the player who can force the other player to stop—by virtue of having moving power—can induce a better outcome than if the other player has moving power (to be defined). But because the Nash equilibrium will never be chosen in either these or the other four games—whichever

---

<sup>9</sup>It is worth noting here that surprise does not arise from incomplete information about player payoffs but rather from incomplete information about the rules of play, or possibly which player has moving power, which will be discussed later.

player possesses moving power—it is not an outcome to which players with a propensity to cycle would be attracted.

Put another way, the building-block concept of stability in noncooperative game theory, that of a Nash equilibrium, tends to cover up the choices that rational players not only would have good reason to make, but have actually made, in the Freedom Game. In particular, the unique Nash equilibrium in this game is very vulnerable: rational players would abandon this equilibrium under two plausible sets of rules, one of which assumes nonmyopic calculations, the other of which assumes cycling to try to outlast an opponent.

Attempting to run an opponent into the ground, so to speak, is hardly an uncommon strategy. The Reagan and Bush administrations pursued this strategy with a vengeance against the Soviet Union, “winning” the Cold War when the Soviet Union collapsed in 1991. Egypt fought five wars with Israel between 1948 and 1973 before both sides, out of despair and futility at fighting again (Israel remained militarily superior, but Egypt had greater manpower), met at Camp David in 1978 and signed a peace treaty in 1979.

When apparently stable and long-lasting relationships rapidly, and often surprisingly, change, we are frequently at a loss to explain why. The instability of the Nash equilibria in the six  $2 \times 2$  games subsumed by the Freedom Game arises from the constraints imposed by the rules of play in standard game theory. When these rules are changed so as to give fuller expression to the range of choices players can and do make, the surprise disappears.

Specifically, these rules, and the rationality calculations underlying them, artificially constrain players. Empirically, these rules are untenable as models of real-world behavior, as the cases previously mentioned illustrate. Philosophically, the situation may be worse: a theory should not proscribe choices that are immediately harmful (e.g., departures from Nash equilibria) but are rationally justifiable in the longer term.

Such departures in the Freedom Game, as I will show for the first set of rules to be analyzed, trigger subsequent departures by the second player that, in the end, may redound to the benefit of both players. Under the second set of rules, in which cycling can occur, the first player may realize a preferred outcome if he or she—with the kind of stamina that moving power provides—is able to force the other player to stop. When this happens, it may be viewed as a less benign force, abridging the free choice of the victimized player.

In sum, I will, using TOM, suggest alternative sets of rules for modeling free choice in games. One set enables players to think ahead and make moves that do not necessarily lead to immediately better outcomes but ultimately improve the lot of one, and sometimes both, players. Another set allows players to repeat choices by cycling, especially when they think their power to hold out against the other player will lead to a better outcome. Indeed, this power may be beneficial, but instead of expanding the free choice of the players, it contracts the choice of the one forced to concede, illustrating how asymmetrical power relationships may repress freedom.

### 3. Order of Play and Unstable Nash Equilibria

The Freedom Game (simultaneous-play version) is shown in Figure 1a,

---

Figure 1a and 1b about here

---

in which R has two strategies,  $s_1$  and  $s_2$ , and C has two strategies,  $t_1$  and  $t_2$ . The payoffs to the players at the resulting four possible outcomes are given by ordered pairs  $(x_1, x_2)$ , where  $x_1$  is the payoff to R and  $x_2$  the payoff to C.

I show in this section under what conditions the unique dominant-strategy Nash equilibrium in the Freedom Game is vulnerable to order-of-play effects. The conditions for dominance, a Nash equilibrium, and instability are the following:

1. *Dominance.* Without loss of generality, assume that  $s_1$  is R's dominant strategy, so

$$a_1 > d_1; \quad b_1 > c_1. \quad (1)$$

That is, whatever strategy C chooses ( $t_1$  or  $t_2$ ), R prefers his payoffs associated with  $s_1$  to those associated with  $s_2$ , making  $s_1$  R's unconditionally better strategy.

2. *Nash equilibrium.* Without loss of generality, assume that  $(b_1, b_2)$  is the unique *Nash equilibrium* that is associated with R's dominant strategy of  $s_1$ , so

$$b_2 > a_2. \quad (2)$$

That is, C prefers her payoff at  $(b_1, b_2)$  to that which she would receive at  $(a_1, a_2)$  if she switched from  $t_2$  to  $t_1$ ; similarly, R prefers  $(b_1, b_2)$  to  $(c_1, c_2)$ —and so would not switch from  $s_1$  to  $s_2$ —because of the dominance of  $s_1$  assumed in (2) above.

3. *Instability.* To ensure that R benefits by moving first and choosing his *dominated strategy* of  $s_2$ , which is unconditionally worse than  $s_1$ , it must be the case that R prefers  $(d_1, d_2)$  to the simultaneous-play Nash equilibrium of  $(b_1, b_2)$ :

$$d_1 > b_1. \quad (3)$$

Also, to ensure that R will induce  $(d_1, d_2)$  rather than  $(c_1, c_2)$  when he chooses  $s_2$ , C must prefer the former outcome to the latter:

$$d_2 > c_2. \quad (4)$$

When inequalities (3) and (4) are both satisfied, the Nash equilibrium,  $(b_1, b_2)$ , is said to be *unstable*.

Combining inequalities (1) and (3) for R, and (2) and (4) for C, gives a complete ordering of payoffs for R and a partial ordering for C:

$$R: a_1 > d_1 > b_1 > c_1; \quad C: b_2 > a_2 \text{ and } d_2 > c_2. \quad (5)$$

The partial ordering for C admits six different complete orderings:

$$\begin{array}{ll} b_2 > a_2 > d_2 > c_2; & d_2 > c_2 > b_2 > a_2; \\ b_2 > d_2 > c_2 > a_2; & d_2 > b_2 > a_2 > c_2; \\ b_2 > d_2 > a_2 > c_2; & d_2 > b_2 > c_2 > a_2. \end{array}$$

That is, these six orderings all preserve the partial ordering of C; because they are the only ones to do so, they are exhaustive.

The single strict ordering of payoffs for R, and the six strict orderings for C, define six strict ordinal games, which are shown in Figure 2. In these

Figure 2 about here

games, 4 indicates a best payoff, 3 next best, 2 next worst, and 1 worst. Thus, the higher the number, the greater the payoff; but because these payoffs are *ordinal*, they indicate only an ordering of outcomes from best to worst, not the degree to which a player prefers one outcome over another.

The payoffs to R and C are given as ordered pairs  $(i, j)$  in parentheses, where  $i$  is the payoff to R and  $j$  the payoff to C. The ordered pairs  $[u, v]$  in brackets just below the parenthetical pairs will be explained in section 4.<sup>10</sup>

I have grouped the six games into the two classes, I and II, mentioned in section 1. In the three class I games, the  $(d_1, d_2)$  lower-left outcome induced when R moves first and chooses his dominated strategy of  $s_2$  is *Pareto-superior*—that is, better for both players—to the upper-right Nash equilibrium of  $(b_1, b_2)$ . In the three class II games, neither  $(d_1, d_2)$  nor  $(b_1, b_2)$  is Pareto-superior to the other outcome.

Before discussing in more detail differences in these two classes of games, I emphasize what they have in common: the Nash equilibrium of  $(b_1, b_2)$  in each class is unstable. Specifically, R has an incentive to choose  $s_2$  first, thereby inducing C—by responding optimally—to choose  $t_1$ , resulting in  $(d_1, d_2)$  rather than the simultaneous-play Nash equilibrium of  $(b_1, b_2)$ .

Interestingly enough, if *both* players have dominant strategies in a  $2 \times 2$  game (e.g., Prisoners' Dilemma), it is never rational for R (or C) to disrupt the unique pure-strategy Nash equilibrium by moving first. To see this, assume, as before, that  $(b_1, b_2)$  is the unique Nash equilibrium. If not only R's strategy associated with this equilibrium is dominant, but C's is too, then in addition to inequality (2),  $c_2 > d_2$ , which reverses inequality (4) that is necessary for instability.<sup>11</sup>

In class I games, not only would R have an incentive to move first, but C would presumably encourage him to do so in order to benefit herself over what she would obtain at the Pareto-inferior Nash equilibrium,  $(b_1, b_2)$ . By comparison, this is emphatically not true of the class II games, in which C does better at  $(b_1, b_2)$ , whereas R does better at  $(d_1, d_2)$ , which he can induce by moving first.<sup>12</sup>

---

<sup>10</sup>The numbers shown above each game are those given in the classification schemes of Brams (1994) and, in parentheses, Rapoport and Guyer (1966).

<sup>11</sup>Specifically, this reversal eliminates R's incentive to move from the Nash equilibrium, because C will not countermove to  $(d_1, d_2)$ .

<sup>12</sup>Felsenthal and Maoz (1987) argue that R must make a self-binding commitment to induce the Pareto-superior  $(d_1, d_2)$  outcome in class I games. But this is unnecessary if the order of play is assumed to be endogenous, because it is in C's interest, too, that R move first. By contrast, a credible commitment is much more called for in class II

If R moves first, the game is, technically, not a 2 x 2 game but a 2 x 4 game, in which C has four strategies—contingent on what R does—rather than two. C's strategies may be described as follows:

$t_i/t_j$ : choose  $t_i$  if R chooses  $s_1$ ,  $t_j$  if R chooses  $s_2$  ( $i, j = 1$  or  $2$ ).

Thus, for example,  $t_2/t_1$  means that C chooses her first strategy ( $t_2$ ) if R chooses his first strategy ( $s_1$ ), and C chooses her second strategy ( $t_1$ ) if R chooses his second strategy ( $s_2$ ).

Assume R moves first in the 2 x 2 Freedom Game of Figure 1a. The 2 x 4 game that results is shown in Figure 1b, with each of the four payoffs in the 2 x 2 game occurring at two different outcomes in the 2 x 4 matrix. For example,  $(a_1, a_2)$  in the 2 x 2 game occurs in the 2 x 4 game when R chooses  $s_1$  and C chooses either  $t_1/t_1$  or  $t_1/t_2$ , because the latter two choices both imply that C chooses  $t_1$  when R chooses  $s_1$  in the 2 x 2 game.

Whereas R has a dominant strategy of  $s_1$  in the 2 x 2 game, he has no dominant strategy in the 2 x 4 game. While R's inequality (5) shows that  $s_1$  is better for him than  $s_2$  when C chooses  $t_1/t_1$ ,  $t_2/t_2$ , or  $t_1/t_2$ ,  $s_2$  is better than  $s_1$  but when C chooses  $t_2/t_1$ . Hence,  $s_1$  is not unconditionally better for R than  $s_2$ . On the other hand,  $t_2/t_1$  for C is *weakly dominant*: it is at least as good, and sometimes better, than any of C's other three strategies—whether R chooses  $s_1$  or  $s_2$ —according to C's inequality (5).

Assume C chooses her dominant strategy in the 2 x 4 Freedom Game in Figure 1b. Anticipating this choice, R's best response is to choose  $s_2$ , resulting in  $(d_1, d_2)$ , the unique dominant-strategy Nash equilibrium in *this* game.<sup>13</sup> In other words, there is a reversal of roles when R moves first: whereas R possesses a dominant strategy in the 2 x 2 game (simultaneous play), C does in the 2 x 4 game (R moves first), leading to a new Nash equilibrium.

The rational choice of  $(d_1, d_2)$  in the 2 x 4 game reinforces my earlier contention: if the order of moves is endogenous, R has an incentive to move

---

games, which Felsenthal and Maoz do not analyze, because of the players' conflict of interest over the order of play.

<sup>13</sup>The outcome  $(b_1, b_2)$  associated with the choice of  $s_1$  by R and  $t_2/t_2$  by C is also a Nash equilibrium. Thus,  $(b_1, b_2)$ 's equilibrium status carries over from the 2 x 2 game, but its selection in the 2 x 4 game is dubious: it would require that C choose a weakly dominated strategy, forsaking his dominant strategy of  $t_2/t_1$ , associated with the  $(d_1, d_2)$  Nash equilibrium, which gives him a higher payoff than  $(b_1, b_2)$  in both class I and class II games.

first. By redefining, in effect, what game is being played, R can induce his preferred outcome of  $(d_1, d_2)$ , because this outcome is the unique dominant-strategy Nash equilibrium in the  $2 \times 4$  game.<sup>14</sup>

C, as I indicated earlier, will have no quarrel with R for moving first in the three class I games in Figure 2, because both players do better at  $(d_1, d_2)$  than  $(b_1, b_2)$ . In the class II games, by comparison, C will be motivated to preempt R: in the resulting  $4 \times 2$  game (i.e., when C moves first, which is not shown),  $(b_1, b_2)$  is the unique dominant-strategy equilibrium—just as it is in the  $2 \times 2$  game—which C prefers to  $(d_1, d_2)$ .

The players face a major problem if each thinks he or she can successfully preempt the other. Then R will choose  $s_2$  when C chooses  $t_2$ , resulting in  $(c_1, c_2)$ . This outcome is worse for *both* players than either  $(b_1, b_2)$  or  $(d_1, d_2)$  in all class II games, wherein precisely this conflict is likely to occur.

This conflict has somewhat the flavor of Chicken, whereby each player does better at his or preferred equilibrium, but each courts disaster if he or she is not successful in implementing it. But if the players recognize the problem of not coordinating their choices in such a “preemption game”—and there is a compromise outcome that benefits both, as in Chicken, even if it is not as good for each as successful preemption—then perhaps the disaster can be averted.<sup>15</sup>

The analysis so far demonstrates how a preemption problem can crop up, even if a simultaneous-play  $2 \times 2$  game has a unique dominant-strategy Nash equilibrium, when the order of moves is endogenous. But this analysis required looking separately at  $2 \times 2$  and  $2 \times 4$  games that had the same payoff structures—only the order of moves was different.

I next describe a dynamic approach to the analysis of the stability of equilibria, in which the  $2 \times 2$  game is assumed to be the only game played, but under two very different sets of rules from those of classical game

---

<sup>14</sup>It is also a Stackelberg equilibrium (von Stackelberg, 1934) for the case in which R is assumed to be the leader—by virtue of moving first—and C the follower. But the first-mover's advantage is eliminated if there is even a slight amount of noise, impairing the follower's ability to detect the leader's prior move (Bagwell, 1995); here I assume this move is perfectly observable.

<sup>15</sup>Game-theoretic models of successful deterrence, based on Chicken, are analyzed in Brams (1985) and Brams and Kilgour (1988); see also Zagare (1987) and Powell (1989).

theory. This approach offers new insight into not only why players might depart from an (unstable) Nash equilibrium but also where they might end up and how they might get there. Thereby TOM makes “surprise” moves both less surprising and more explicable in terms of the dynamics of rational choice.<sup>16</sup>

#### 4. The Theory of Moves (TOM)

The starting point of TOM is a payoff matrix, or *configuration*, in which the order of play is not specified.<sup>17</sup> In fact, players are assumed not even to choose strategies but instead to move and countermove from outcomes, or states, by looking ahead and using “backward induction” to determine the rationality of both their moves and those of an opponent.

Because game theory assumes that players choose strategies simultaneously,<sup>18</sup> it does not raise questions about the rationality of moving or departing from outcomes—at least beyond an immediate departure, à la Nash. In fact, however, most real-life games do not start with simultaneous strategy choices but commence at outcomes. The question then becomes whether a player, by departing from an outcome, can do better not just in an immediate or myopic sense but, rather, in an extended or nonmyopic sense.

In the case of 2 x 2 games, TOM postulates four *rules of play*, which describe the possible choices of the players at different stages:

1. Play starts at an outcome, called the *initial state*, which is at the intersection of the row and column of a 2 x 2 payoff matrix.
2. Either player can unilaterally switch its strategy, and thereby change the initial state into a new state, in the same row or

---

<sup>16</sup>For a more general argument, and plethora of examples, that “surprise” depends on the rules of play—specifically, not understanding them or their consequences—see Casti (1994).

<sup>17</sup>The subsequent description of TOM, which also encompasses rules not discussed here or in later sections that allow for threats, misperception, and the like, is adapted from Brams (1994).

<sup>18</sup>Strategies may allow for sequential choices, as I showed in section 2, but the classical theory does not make endogenous who moves first, as TOM does, but instead specifies a fixed order of play (simultaneous or sequential).

column as the initial state.<sup>19</sup> The player who switches, who may be either R or C, is called player 1 (P1).

3. Player 2 (P2) can respond by unilaterally switching its strategy, thereby moving the game to a new state.
4. The alternating responses continue until the player (P1 or P2) whose turn it is to move next chooses not to switch its strategy. When this happens, the game terminates in a *final state*, which is the *outcome* of the game.

Note that the sequence of moves and countermoves is strictly alternating: first, say, R moves, then C moves, and so on, until one player stops, at which point the state reached is final and, therefore, the outcome of the game.<sup>20</sup>

The use of the word “state” is meant to convey the temporary nature of an outcome, before players decide to stop switching strategies. I assume that no payoffs accrue to players from being in a state unless it is the final state and, therefore, becomes the outcome (which could be the initial state if the players choose not to move from it).

To assume otherwise would require that payoffs be cardinal rather than ordinal, with players accumulating them as they pass through states. I eschew this assumption in part because I think payoffs to players in most real-life games cannot be quantified and summed across the states visited. More significant, payoffs in the games that most interest me depend overwhelmingly on the final state reached, not on how it was reached. In politics, for example, the payoff for most politicians is not in campaigning, which is arduous and costly, but in winning.

Rule 1 differs radically from the corresponding rule of play in classical game theory, in which players simultaneously choose strategies in a matrix game that determines an outcome. Instead of starting with strategy choices,

---

<sup>19</sup>I do not use “strategy” in the usual sense to mean a complete plan of responses by the players to all possible contingencies allowed by rules 2-4, because this would make the normal form unduly complicated to analyze. Rather, *strategies* refer to the choices made by players that define a state, and *moves and countermoves* to their subsequent strategy switches from an initial state to a final state in an extensive-form game, as allowed by rules 2-4. For another approach to combining the normal and extensive forms, see Mailath, Samuelson, and Swinkels (1993).

<sup>20</sup>An emendation in the rules of TOM that allows for backtracking would be appropriate in games of incomplete information, wherein players make mistakes that they wish to rectify. For more on possible rules changes under TOM, see Brams (1994).

I assume that players are already in some state at the start of play and receive payoffs from this state if they stay. Based on these payoffs, they decide, individually, whether or not to change this state in order to try to do better.<sup>21</sup>

To be sure, some decisions are made collectively by players, in which case it would be reasonable to say that they choose strategies from scratch, either simultaneously or by coordinating their choices. But if, say, two countries are coordinating their choices, as when they agree to sign a treaty, the most important strategic question is what individualistic calculations led them to this point. The formality of jointly signing the treaty is the culmination of their negotiations, which covers up the move-countermove process that preceded it. This is precisely what TOM is designed to uncover.

In summary, play of a game starts in a state, at which players accrue payoffs only if they remain in that state so that it becomes the outcome of the game. If they do not remain, they still know what payoffs they would have accrued had they stayed; hence, they can make a rational calculation of the advantages of staying versus moving. They move precisely because they calculate that they can do better by switching states, anticipating a better outcome if and when the move-countermove process finally comes to rest.

Rules 1–4 say nothing about what causes a game to end, but only when: termination occurs when a “player whose turn it is to move next chooses not to switch its strategy” (rule 4). But when is it rational not to continue moving, or not to move in the first place from the initial state?

To answer this question, I posit a rule of *rational termination* (first proposed in Brams, 1983, pp. 106-107), which has been called “inertia” by Kilgour and Zagare (1987, p. 94). It prohibits a player from moving from an initial state unless doing so leads to a better (not just the same) final state:

5. A player will not move from an initial state if this move
  - (i) leads to a less preferred final state (i.e., outcome); or
  - (ii) returns play to the initial state (i.e., makes the initial state the outcome).

---

<sup>21</sup>Alternatively, players may be thought of as choosing strategies initially, after which they perform a thought experiment of where moves will carry them once a state is selected. The concept of an “anticipation game,” developed later, advances this idea, which might be considered dynamic thinking about the static play of a matrix game. Generally, however, I assume that “moves” describe actions, not just thoughts, though I readily admit the possibility of the thought interpretation.

I will discuss shortly how rational players, starting from some initial state, determine, by using backward induction, what the outcome will be.

Condition (i) of rule 5, which precludes moves that result in an inferior state, needs no defense. But condition (ii), which precludes moves to the same state because of cycling back to the initial state, is worth some elaboration. It says that if it is rational, after P1 moves, for play of the game to cycle back to the initial state, P1 will not move in the first place. After all, what is the point of initiating the move-countermove process if play simply returns to “square one,” given that the players receive no payoffs along the way (i.e., before an outcome is reached)?

Not only is there no gain from cycling, but in fact there may be a loss because of so-called transaction costs that players suffer by virtue of making moves that, ultimately, do not change the situation. Therefore, it seems sensible to assume that P1 will not trigger a move-countermove process if it only returns the players to the initial state, making it the outcome.

I call rule 5 a *rationality rule*, because it provides the basis for players to determine whether they can do better by moving from a state or remaining in it. Still another rationality rule is needed to ensure that both players take into account each other’s calculations before deciding to move from the initial state. I call this rule the *two-sidedness rule*:

6. Given that players have complete information about each other’s preferences and act according to the rules of TOM, each takes into account the consequences of the other player’s rational choices, as well as its own, in deciding whether to move from the initial state or subsequently, based on backward induction. If it is rational for one player to move and the other player not to move from the initial state, then the player who moves takes *precedence*: its move overrides the player who stays, so the outcome is that induced by the player who moves.

Because players have complete information, they can look ahead and anticipate the consequences of their moves. I next show how, using backward induction, they do this. In the process, I link order-of-moves effects, due to unstable equilibria (section 3), to the results given by TOM, which predicts the moves that players will make from these equilibria.

## 5. Unstable Nash Equilibria and TOM

I begin by analyzing where the players will end up when the initial state is the unique dominant-strategy Nash equilibrium of  $(b_1, b_2)$  in the  $2 \times 2$  Freedom Game (Figure 1a).<sup>22</sup>

**Theorem 1.** *In the  $2 \times 2$  Freedom Game,  $(d_1, d_2)$  is the outcome to which the players will move, according to TOM, from the unique dominant-strategy Nash equilibrium of  $(b_1, b_2)$ .*

**Proof.** There are two cases to consider:

### *Case 1: R Acts First*

If R acts first, the counterclockwise progression of moves from  $(b_1, b_2)$  back to  $(b_1, b_2)$ —with the player (R or C) who makes the next move shown below each state in the alternating sequence—is as follows (see Figure 1a):

State 1		State 2		State 3		State 4		State 1
R		C		R		C		
<i>R starts:</i> $(b_1, b_2)$	→	$(c_1, c_2)$	→	<del><math>(d_1, d_2)</math></del>	→	$(a_1, a_2)$	→	$(b_1, b_2)$
<i>Survivor:</i> $(d_1, d_2)$		$(d_1, d_2)$		$(d_1, d_2)$		$(b_1, b_2)$		

The *survivor* is determined by working backward, after a putative cycle has been completed.

Assume the players' alternating moves have taken them counterclockwise from  $(b_1, b_2)$  to  $(c_1, c_2)$  to  $(d_1, d_2)$  to  $(a_1, a_2)$ , at which point C must decide whether to stop at  $(a_1, a_2)$  in state 4 or complete the cycle and return to  $(b_1, b_2)$ . Given inequality (2), C prefers  $(b_1, b_2)$  to  $(a_1, a_2)$ , so  $(b_1, b_2)$  is listed as the survivor below  $(a_1, a_2)$ : because C would move the process back to  $(b_1, b_2)$  should it reach  $(a_1, a_2)$ , the players know that if the move-countermove process reaches state 4, the outcome will be  $(b_1, b_2)$ .

Knowing this, would R at state 3 move to  $(a_1, a_2)$ ? Given inequality (3), R prefers  $(d_1, d_2)$  to the survivor at state 3—namely,  $(b_1, b_2)$ —so the answer is no. Hence,  $(d_1, d_2)$  becomes the survivor when R must choose between stopping at  $(d_1, d_2)$  and moving to  $(a_1, a_2)$  at state 4—which, as I have just shown, would become  $(b_1, b_2)$  once  $(a_1, a_2)$  is reached.

---

<sup>22</sup>Where the players end up depends on the endstate, or *anchor*, from which the backward induction proceeds, which I assume here—for reasons given in section 4—is after one complete cycle. This assumption defines a finite extensive-form game, or game tree, to which I apply backward induction.

At state 2, C prefers moving to  $(d_1, d_2)$  than stopping at  $(c_1, c_2)$  by inequality (4). Consequently,  $(d_1, d_2)$  again is the survivor if the process reaches  $(c_1, c_2)$  at state 2. Similarly, at state 1 R prefers the previous survivor,  $(d_1, d_2)$ , to  $(b_1, b_2)$  from inequality (3), so  $(d_1, d_2)$  is the survivor at this state as well.

The fact that  $(d_1, d_2)$  is the survivor at state 1 means that it is rational for R initially to move to  $(c_1, c_2)$ , and C subsequently to move to  $(d_1, d_2)$ . At  $(d_1, d_2)$  the process will stop, making  $(d_1, d_2)$  the rational choice if R has the opportunity to move first from state 1. That is, after working *backward* from C's choice of completing the cycle or not at state 4, the players can reverse the process and, looking *forward*, determine that it is rational for R to move from  $(b_1, b_2)$  to  $(c_1, c_2)$ , and C to move from  $(c_1, c_2)$  to  $(d_1, d_2)$ , at which point R will stop the move-countermove process at  $(d_1, d_2)$ .

Notice that R does better at  $(d_1, d_2)$  than at  $(b_1, b_2)$ , where it could have terminated play at the outset, and C does better at  $(d_1, d_2)$  than at  $(c_1, c_2)$ , where it could have terminated play, given that R moves first. I indicate that  $(d_1, d_2)$  is the consequence of backward induction by underscoring this state in the progression; it is the state at which *stoppage* of the process occurs, which I indicate by the vertical line blocking the arrow emanating from  $(d_1, d_2)$ .

### Case 2: C Acts First

If C acts first, the progression of moves from  $(b_1, b_2)$  back to  $(b_1, b_2)$  is clockwise. But this time backward induction seems to leave open which state will be the final survivor. In particular, given that R chooses to complete the cycle at state 4 because  $b_1 > c_1$  of inequality (1)—making  $(b_1, b_2)$  the survivor at state 4—C's inequality (5) provides no information about which choice C will make at state 3—stay at  $(d_1, d_2)$ , or move to  $(c_1, c_2)$  and thence to  $(b_1, b_2)$ :

	State 1		State 2		State 3		State 4		State 1
	C		R		C		R		
<i>C starts:</i>	$(b_1, b_2)$	?	$(a_1, a_2)$	?	$(d_1, d_2)$	?	$(c_1, c_2)$	→	$(b_1, b_2)$
<i>Survivor:</i>							$(b_1, b_2)$		

This indeterminacy disappears, however, when the consequences of two possibilities at state 3—either  $(b_1, b_2)$  survives or  $(d_1, d_2)$  survives, depending on whether  $b_2 > d_2$  or  $d_2 > b_2$ —are analyzed:

- $(b_1, b_2)$  survives. Then R will choose  $(a_1, a_2)$  at state 2 because of R's inequality (5), and C will choose  $(b_1, b_2)$  at state 1 because of inequality (2), so  $(b_1, b_2)$  will be the survivor at state 1.
- $(d_1, d_2)$  survives. Then R will choose  $(a_1, a_2)$  at state 2 because  $a_1 > d_1$  of inequality (1), and C will choose  $(b_1, b_2)$  at state 1 because of inequality (2), so  $(b_1, b_2)$  will be the survivor at state 1.

In either case,  $(b_1, b_2)$  will be the outcome when C acts first, which is to say that C will not move from this state. But because I showed in case 1 that R will move from this state and induce  $(d_1, d_2)$ ,  $(d_1, d_2)$  will be the outcome because of the precedence of moving over staying (rule 6 in section 4).  $\square$

Thus, the unique Nash equilibrium of  $(b_1, b_2)$  in the Figure 1a game is not only unstable (section 3) but also one from which the players would depart and, according to TOM, move to  $(d_1, d_2)$ . I next consider whether there are any additional  $2 \times 2$  strict ordinal games, apart from the six given in Figure 2, that have unique dominant-strategy Nash equilibria from which the players would depart.

**Theorem 2.** *The six games in Figure 2 are the only  $2 \times 2$  strict ordinal games in which the players will move from a unique dominant-strategy Nash equilibrium. Consequently, these games are the same as those in which order of play, due to instability, matters (analyzed in section 3).*

**Proof.** To show that there are no other games among the 78 in which players would move from a dominant-strategy Nash equilibrium, one must show that if either of the necessary conditions for instability given in section 2 fails—inequality (3) or inequality (4)—the backward induction used in the proof of Theorem 1 will lead to the players' staying at  $(b_1, b_2)$  rather than moving to  $(d_1, d_2)$ . Consider, first, case 1 when R acts first and the possible failure of each instability condition:

- Inequality (3) is reversed, so  $b_1 > d_1$ . Then the final survivor at state 1 will be  $(b_1, b_2)$  rather than  $(d_1, d_2)$ .
- Inequality (4) is reversed, so  $c_2 > d_2$ . Then the next-to-final survivor at state 2 will be  $(c_1, c_2)$ , which will be displaced as final survivor by  $(b_1, b_2)$  at state 1 since  $b_1 > c_1$  by inequality (1).

Hence, R will not move from  $(b_1, b_2)$  if either inequality (3) or inequality (4) is reversed.

As for C, it is easy to show that the reversal of either inequality does not change the backward-induction analysis in case 2, so C, as before, will not move from  $(b_1, b_2)$  with the reversal of either inequality. Therefore, only the satisfaction of both inequalities induces the players to move from  $(b_1, b_2)$ , which are exactly the two conditions that produce the order-of-play effects (i.e., make the dominant-strategy Nash equilibrium unstable).  $\square$

Taken together, Theorems 1 and 2 demonstrate that order of play matters in  $2 \times 2$  games with unique dominant-strategy Nash equilibria if and only if the players would, according to TOM, move from these equilibria to another state. The state that they would move to is the unique Nash equilibrium in the  $2 \times 4$  game in which R (i.e., the player with the dominant strategy in the  $2 \times 2$  game) moves first and chooses his dominated strategy.

For the six games in Figure 2, I indicate that the upper-right  $(b_1, b_2)$  will go into the lower-left  $(d_1, d_2)$  by putting  $(d_1, d_2)$  in brackets, just below  $(b_1, b_2)$ , in each payoff matrix. In game 27, for example,  $[3,4]$  is shown below  $(2,3)$  in the upper-right cell.

The outcome into which a state goes is called the *nonmyopic equilibrium* (NME) from that state. NMEs may be viewed as the consequence of both players' looking ahead and making rational calculations of where the move-countermove process will transport them, based on the rules of TOM, from each of the four possible initial states.

Backward-induction analysis from each state in game 27 shows that each state will go into  $(3,4)$ . Thus, wherever play starts, the players can anticipate that they will end up at  $(3,4)$ , making it the unique NME in game 27. This is also true of  $(3,4)$  in game 28, but not in game 48, the third class I game. Starting at  $(4,2)$  in game 48, the players will not depart from this state, making  $(4,2)$  as well as  $(3,4)$  an NME in this game.

Like game 48, all games in class II contain at least two NMEs. But some of these NMEs are *indeterminate*, because there is a conflict over who will move first. In game 50, for example, if  $(1,1)$  is the initial state,  $[2,4]/[4,3]$  indicates that when R moves first from  $(1,1)$ ,  $(2,4)$  will be outcome, whereas when C moves first,  $(4,3)$  will be the outcome.<sup>23</sup>

---

<sup>23</sup>Actually, the result of backward induction by R from  $(1,1)$  in game 50 is  $(3,2)$  rather than  $(4,3)$ . But, as I argue in Brams (1994, p. 115, fn. 20), the players would have a common interest in implementing the Pareto-superior  $(4,3)$  to  $(3,2)$  when there is clockwise movement from  $(1,1)$ . However, the implementation of  $(4,3)$  would require a binding commitment on the part of R not to move on from  $(4,3)$  to  $(2,4)$ , which is not

Because R prefers (4,3) whereas C prefers (2,4), each player will try to hold out longer in order to induce the other player to move first. Who wins in this struggle will depend on which player has “order power”—that is, who can determine the order of moves, starting at (1,1) (Brams, 1994, ch. 5).

Every  $2 \times 2$  game contains at least one NME, because from each initial state there is an outcome (perhaps indeterminate) of the move-countermove process. If this outcome is both determinate and the same from every initial state, then it is the only NME; otherwise, there is more than one NME.

In game 56 in Figure 2, there are three different NMEs, which is the maximum number that can occur in a  $2 \times 2$  strict ordinal game; the minimum, as already noted, is one. All except two of the 78  $2 \times 2$  games (game 56 and Chicken) have either one or two NMEs.

The four bracketed states of each game in Figure 2 define what I call the *anticipation matrix*, with each state in this matrix an NME. Insofar as players choose strategies as if they were playing a game based on this matrix, one can determine which NMEs are Nash equilibria in *this* game and therefore likely to be chosen.<sup>24</sup>

To summarize, where players start in the original games in Figure 2—including the unique dominant-strategy Nash equilibria—may not be where they end up, according to TOM. Thus, an original game may mask a good deal of instability when the players can move and countermove from states.<sup>25</sup> And it is this instability that induces players to make choices that, while

---

assumed possible in noncooperative game theory. I conclude: “I do not see an airtight case being made for either (3,2) or (4,3) as *the* NME from (1,1) when Column moves first, which nicely illustrates the nuances that TOM surfaces that the rules of standard game theory keep well submerged.” Incidentally, game 50 is the only game of the 78  $2 \times 2$  strict ordinal games in which this kind of ambiguity about NMEs arises.

<sup>24</sup>Because the NMEs in games 27 and 28 are all the same, the strategies of the players in their anticipation games are indistinguishable, making all four states Nash equilibria. In the four other games,  $s_1$  is R’s (weakly) dominant strategy; C’s best response leads to a unique dominant-strategy Nash equilibrium in each anticipation game. Only in game 49 is the Nash equilibrium in the anticipation game, [2,4], also the Nash equilibrium, (2,4), in the original game.

<sup>25</sup>Even the choice of a dominant-strategy Nash equilibrium in the anticipation games offers no assurance that players will stay at this state. Indeed, except for one of the two [4,3] Nash equilibria in the anticipation game of game 50, the players will move from *every* Nash equilibrium in the six anticipation games to some different NME.

disallowed by the rules of standard game theory, are allowed by rules 2-4 of TOM.

## 6. Cyclic Games and Moving Power

In this section, I postulate two new rationality rules (rules 5' and 6' instead of rules 5 and 6, given in section 4); rule 5' permits players to cycle in a matrix, and rule 6' enables one player, if he or she possesses "moving power," to terminate the cycling.

Recall from section 4 that the rules of play (i.e., rules 1-4) say nothing about what causes a game to end, only when. Rule 5, which forbids moves that do not lead to a better outcome and thus precludes cycling back to the initial state, provides one answer. But this ban on cycling may not be realistic, as many protracted conflicts (e.g., the Arab-Israeli conflict)—in which the protagonists have revisited the past again and again—make unmistakably clear.

To try to capture the cyclic aspect of certain conflicts, and give players the ability to make choices in which they repeat themselves (why they may want to do so will be considered shortly), I next define a class of games in which cycling is possible by precluding a class of games in which it is not. Rule 5' provides a sufficient condition for cycling not to occur:

- 5'. If, at any state in the move-countermove process, a player whose turn it is to move next receives its best payoff (i.e., 4), it will not move from this state.

Rule 5', in fact, precludes cycling in 42 of the 78  $2 \times 2$  games, 21 of which contain a mutually best (4,4) state. Of the remaining 36 games, they fall into three classes: strongly cyclic (9 games), moderately cyclic (18 games), and weakly cyclic (9 games).

The six  $2 \times 2$  games subsumed by the Freedom Game are all moderately cyclic. Before illustrating this kind of cyclicity, consider the circumstances under which players, who know not only their own payoffs but also have complete information about the payoffs of their opponents, would have an incentive to cycle to try to outlast an opponent.

By "outlasting" an opponent, I mean that one (stronger) player can force the other (weaker) player to stop the move-countermove process at a state where the weaker player has the next move. Forcing stoppage at such a state, however, may not always lead to an outcome favorable to the stronger

player. In fact, a stoppage-forcing strategy is unproductive in four of the six  $2 \times 2$  games that the Surprise Game embraces, leading to a worse outcome than if the stronger player had stopped the cycling rather than forcing the weaker player to do so.

Rule 5' specified what players would not do—namely, move from a best (4) state when it was their turn to move. By precluding cycling in 42 of the 78 games, however, this rule did not say anything about where cycling would stop in the remaining 36 cyclic games. In these games, a new rationality rule is needed to say at what, if any, state play will terminate:

- 6'. In a cyclic game of complete information, player 1 (P1) will move from an initial state, even if play returns to this state and repeatedly cycles, if P1
- (i) has “moving power”; and
  - (ii) can induce a better outcome with this power.

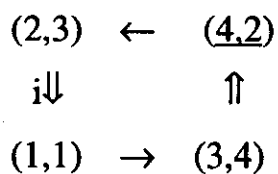
P1 has *moving power* if he or she can induce player 2 (P2) eventually to stop, in the process of cycling, at one of the two states at which P2 has the next move. The state at which P2 stops is that which P2 prefers (Brams, 1995, ch. 4).

Define moving power to be *effective* if the outcome that each player can induce with this power is better for it than the outcome that the other player can induce. Given this definition, we have

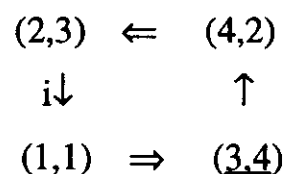
**Theorem 3.** *Moving power is effective in two of the six games in Figure 2 (games 48 and 56).*

I will illustrate this result for game 48 (the reader can verify it for game 56) and later illustrate why moving power is not effective in game 49 (the reader can verify this for games 27, 28, and 50). The arrows shown below (ignore for now the distinction between the single and double arrows) illustrate the

*R can induce (4,2) in game 48*



*C can induce (3,4) in game 48*

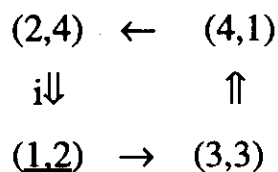


the cyclicity of game 48 in a counterclockwise direction: starting at the upper right state, C benefits moving from (4,2) to (2,3); R does not benefit moving from (2,3) to (1,1)—that is why “i” (for impediment) is next to this arrow—but departs from a 2, not a 4, state; C benefits moving from (1,1) to (3,4); and R benefits moving from (3,4) to (4,2). Because no player, when it is his or her turn to move, ever departs from his or her best (i.e., 4) state, and because game 48 has one impediment, this game is moderately cyclic (games with no impediments are strongly cyclic, and games with two impediments are weakly cyclic).<sup>26</sup>

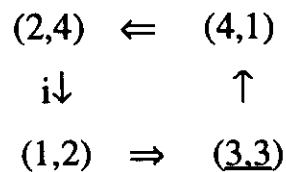
To show what outcome R can induce if he has moving power, which might be thought of as greater endurance, let his moves (vertical, as illustrated on the left side of the above diagram) be represented by double arrows. C, whose (horizontal) moves are represented by single arrows, must stop in the cycling at either (1,1) or (4,2), from where her single arrows emanate that indicate she has the next move. Since she would prefer to stop at (4,2) than (1,1), R can induce his best outcome if he has moving power. On the other hand, if C has moving power (right side of the above diagram), she can force R to stop at either (2,3) or (3,4), from where his single arrows emanate that indicate he has the next move. Since he would prefer to stop at (3,4) rather than (2,3), C can induce her best outcome if she has moving power. Thus, the possession of moving power benefits the player who possesses it—compared with the other player’s possession of it—so it is effective in game 48.

This is not the case in game 49. Applying the foregoing reasoning to this game, we see that R can induce only (1,2)—C prefers this to (4,1), the

*R can induce (1,2) in game 49*



*C can induce (3,3) in game 49*



other state where she moves from—but C can induce (3,3)—R prefers this to (2,4), the other state he moves from. But since R also prefers (3,3) to (1,2),

---

<sup>26</sup>It is shown in Brams (1994, pp. 90-91) that the 78 2 x 2 strict ordinal games can be cyclic in at most one direction (either clockwise or counterclockwise). Furthermore, none of the 12 symmetric games (e.g., Chicken and Prisoners’ Dilemma) is cyclic, because their noncyclicity in one direction implies it in the other direction.

moving power is not effective: R cannot induce a better outcome when he has it than when C has it. Rather, it is “irrelevant” in game 49, because it would be in R’s interest to stop at (3,3), even if he has moving power, rather than to force C to stop at (1,2).<sup>27</sup> In Figure 2, the outcomes each player can induce with moving power are indicated by superscripts *r* (for row) and *c* (for column), with an outcome starred if moving power is irrelevant and *both* players would find it rational for play to stop at this outcome.

Moving power may be thought of as a constraint on free choice if the player who possesses it (i) can force the other player to stop and (ii) has good reason to do so. As we have seen in four of the six games that the Freedom Game subsumes, however, the possessor has no reason to exercise this kind of power—preferring to stop himself or herself in the move-countermove process—which renders its possession irrelevant. Hence, even if there is a power asymmetry in such games, it does not translate into an effective constraint. On the other hand, in the two games in which moving power is effective, the possession of this power by one player will prevent the other player from obtaining a preferred outcome.

One lesson in all this is that power asymmetries sometimes may stifle free choice, but such a consequence is far from inevitable. In four of the games under consideration, moving power is irrelevant. Nonetheless, other kinds of power based on TOM are effective in two games in which moving power is not (“threat power” in games 49 and 50; “order power” is effective in game 56, along with moving power). Thus, a second lesson is that free choice is also sensitive to what kinds of power can be exercised. When these different kinds are taken into account, only two of the six games (27 and 28) are unaffected by one player’s possession of power (Brams, 1994).

Although one player’s possession of power in these two games does not help it vis-à-vis the other, both players benefit from the exercise of either moving or (deterrent) threat power. This is because either R or C, with this power, can induce (3,4), which is Pareto-superior to the Nash equilibrium of (2,3) in game 27 and the Nash equilibrium of (2,2) in game 28. Hence, the exercise of power in these games has a salutary effect, reinforcing the choice of the unique NME (albeit under different rules).

---

<sup>27</sup>More generally, moving power is *irrelevant* when the outcome induced by one player is better for both. There is still a third category of moving power, which I call *ineffective*—each player prefers the outcome that the other player can induce with it to that which he or she can induce—but it does not apply to any of the six games in Figure 2 (it is applicable only to 4 of the 78 2 x 2 strict ordinal games)

## 7. Conclusions

In this paper I showed that in both standard game theory and TOM, the incentive to move from a unique dominant-strategy Nash equilibrium in the generic Freedom Game, which subsumes six specific  $2 \times 2$  strict ordinal games, has a rational-choice justification. However, the justification in the standard theory is awkward. It requires comparing the Nash equilibrium in the  $2 \times 2$  game with the Nash equilibrium that R can induce if he moves first in a  $2 \times 4$  game. If the equilibrium in the  $2 \times 4$  game is better for R, then R would have an incentive to move first, making the equilibrium in the  $2 \times 2$  game unstable.

TOM, by contrast, does not presume different game forms or a first-moving player. Rather, play starts at an outcome, or state, with the order of moves endogenous. Players move and countermove according to rules of play 1-4; under the first set of rationality rules (5 and 6), they look ahead and are able to anticipate ending up at NMEs from each state.

It turns out that the six  $2 \times 2$  games in which players would move from the dominant-strategy Nash equilibrium to a different state (i.e., an NME) are exactly the  $2 \times 2$  games with unstable Nash equilibria. What TOM offers that the standard theory does not is a dynamic and parsimonious rationale for the players' making their moves and countermoves, which were used to model two empirical cases in Brams (1995).

Specifically, I showed that the surprises of Pearl Harbor (modeled by game 56) and the Sadat peace initiative (modeled by game 27) stem from thinking that the player with the dominant strategy (Japan and Egypt, respectively) would have no incentive to depart from this strategy. But the outcome induced by the other player's (the United States' and Israel's, respectively) best response to these dominant strategy choices—that is, the unique Nash equilibrium in these games—has no equilibrium status, according to TOM, when play starts in this state.

The migration of the players from the Nash equilibrium in these cases to a different NME, I argued, is therefore not surprising. The fact that there was so much surprise at the time led me to postulate that, even if the games played were games of complete information, there was incomplete information about either the rules of play or the rationality rules, at least on the part of the surprised player.

Because different rules may generate radically different equilibrium results, it is important to try to ascertain those that are most plausible in a

situation. Rules 1-6, in my opinion, are eminently plausible in many situations. Players often do think more than one step ahead and calculate, on this basis, a series of moves and countermoves that they anticipate they and an opponent will make.

A theory should not deny players the opportunity to exercise fully their freedom of choice—and a more far-sighted rationality, based on this freedom to move and countermove, according to rules 1-4. This may entail not just the kind of nonmyopic thinking of rules 5 and 6, which precludes cycling, but also the possibility of cycling, as governed by rules 5' and 6', that occurred in the superpower rivalry and continues in the Arab-Israeli conflict.

In the case of cycling, I showed that moving power is effective in two of the six games subsumed by the Freedom Game (48 and 56), making it rational for the player who possesses it to try to outlast the other player by restricting his or her freedom of choice. In the other four games, moving power is irrelevant, making it in the interest of both players to stop at the same outcome.

In the latter games—as in the first two games—the outcome induced by moving power is never the Nash equilibrium. In two of these games (27 and 28) it is Pareto-superior to it, and in the other two (49 and 50) it benefits one player at the expense of the other. Thus, even when moving power is irrelevant in cyclic games, one would expect it to be exercised by at least one player to try to undermine the Nash equilibrium.

To conclude, the rules of standard game theory fail, on occasion, to capture the more long-range choices that players can and do make in the Freedom Game, especially departures from its unique Nash equilibrium. The first set of new rules expands player choices by assuming that players have the ability to think ahead in plotting their moves, but it prohibits cycling. The second set allows for cycling, affording the players still more choices, but these may be limited by the exercise of moving power. Whichever set is most appropriate to model a given situation, each allows players a freedom to make moves that is constrained only by their rationality and the possibly greater moving power of an opponent.

## References

- Bagwell, Kyle (1995). "Commitment and Observability in Games." *Games and Economic Behavior* 8, no. 2 (February): 271-280.
- Brams, Steven J. (1983). *Superior Beings: If They Exist, How Would We Know? Game-Theoretic Implications of Omniscience, Omnipotence, Immortality, and Incomprehensibility*. New York: Springer-Verlag.
- Brams, Steven J. (1985). *Superpower Games: Applying Game Theory to Superpower Conflict*. New Haven, CT: Yale University Press.
- Brams, Steven J. (1993). "Theory of Moves." *American Scientist* 81, no. 6 (December): 562-570.
- Brams, Steven J. (1994). *Theory of Moves*. Cambridge, UK: Cambridge University Press.
- Brams, Steven J. (1995). "The Rationality of Surprise: Unstable Nash Equilibria and the Theory of Moves." Preprint, Department of Politics, New York University.
- Brams, Steven J., and D. Marc Kilgour (1988). *Game Theory and National Security*. New York: Basil Blackwell.
- Brams, Steven J., and Walter Mattli (1993). "Theory of Moves: Overview and Examples." *Conflict Management and Peace Science* 12, no. 2 (Spring): 1-39.
- Casti, John L. (1994). *Complexification: Explaining a Paradoxical World Through the Science of Surprise*. New York: HarperCollins.
- Hamilton, Jonathan H., and Steven M. Slutsky (1990). "Endogenous Timing in Duopoly Games: Stackelberg or Cournot Equilibria." *Games and Economic Behavior* 2, no. 1 (March): 29-46.
- Hamilton, Jonathan H., and Steven M. Slutsky (1993). "Endogenizing the Order of Moves in Matrix Games." *Theory and Decision* 34, no. 1 (January): 47-62.
- Hirshleifer, Jack (1985). "Protocol, Payoff, and Equilibrium: Game Theory and Social Modelling." Working Paper #366, Department of Economics, University of California, Los Angeles (March).

- Kilgour, D. Marc, and Frank C. Zagare (1987). "Holding Power in Sequential Games." *International Interactions* 13, no. 2: 91-114.
- Mailath, George J., Larry Samuelson, and Jeroen Swinkels (1993). "Extensive Form Reasoning in Normal Form Games." *Econometrica* 61, no. 2 (March): 273-302.
- Maoz, Zeev, and Dan S. Felsenthal (1987). "Self-Binding Commitments, the Inducement of Trust, Social Choice, and the Theory of International Cooperation." *International Studies Quarterly* 31, no. 2 (June): 177-200.
- Powell, Robert (1989). *Nuclear Deterrence Theory: The Search for Credibility*. Cambridge, UK: Cambridge University Press.
- Rapoport, Anatol, and Melvin J. Guyer (1966). "A Taxonomy of 2 x 2 Games." *General Systems: Yearbook of the Society for General Systems Research* 11: 203-214.
- Rosenthal, Robert W. (1991). "A Note on Robustness of Equilibria with Respect to Commitment Opportunities." *Games and Economic Behavior* 3, no. 2 (May): 237-243.
- van Damme, Eric (1991). *Stability and Perfection of Nash Equilibria*. Heidelberg, Germany: Springer-Verlag.
- van Damme, E., and S. Hurkens (1993). "Commitment Robust Equilibria and Endogenous Timing." Discussion Paper No. 9356, Center for Economic Research, Tilburg University, The Netherlands (July).
- von Stackelberg, Henrich (1934). *Marktform und Gleichgewicht*. Berlin: J. Springer. Trans. by Alan Peacock as *The Theory of Market Economy*. London: William Hodge, 1952.
- Zagare, Frank C. (1987). *The Dynamics of Deterrence*. Chicago: University of Chicago Press.

**Figure 1a. Freedom Game (Simultaneous Play)**

		<i>Column (C)</i>		
		$t_1$	$t_2$	
<i>Row (R)</i>	$s_1$	$(a_1, a_2)$	<u><math>(b_1, b_2)</math></u>	← Dominant strategy
	$s_2$	$(d_1, d_2)$	$(c_1, c_2)$	

**Figure 1b. Freedom Game (R Moves First)**

		<i>Column (C)</i>				
		$t_1/t_1$	$t_2/t_2$	$t_1/t_2$	$t_2/t_1$	
<i>Row (R)</i>	$s_1$	$(a_1, a_2)$	<u><math>(b_1, b_2)</math></u>	$(a_1, a_2)$	$(b_1, b_2)$	
	$s_2$	$(d_1, d_2)$	$(c_1, c_2)$	$(c_1, c_2)$	<u><math>(d_1, d_2)</math></u>	↑ Dominant strategy

**Key:**  $(x_1, x_2) = (\text{payoff to R, payoff to C})$

$t_i/t_j$ : choose  $t_i$  if R chooses  $s_1$ ,  $t_j$  if R chooses  $s_2$  ( $i, j = 1$  or  $2$ ).

Nash equilibria underscored

Complete ordering for R:  $a_1 > d_1 > b_1 > c_1$

Partial ordering for C:  $b_2 > a_2$  and  $d_2 > c_2$

**Figure 2. Six 2 x 2 Strict Ordinal Games Subsumed by Freedom Game (Simultaneous Play)**

*Case I:  $(d_1, d_2)$  Pareto-superior to Nash equilibrium of  $(b_1, b_2)$*

27 (47)	28 (48)	48 (57)																								
<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">(4,1)</td> <td style="padding: 2px;">(2,3)</td> </tr> <tr> <td style="padding: 2px;">[3,4]</td> <td style="padding: 2px;">[3,4]</td> </tr> <tr> <td style="padding: 2px;"><u>(3,4)<sup>c*</sup></u></td> <td style="padding: 2px;">(1,2)<sup>r</sup></td> </tr> <tr> <td style="padding: 2px;">[3,4]</td> <td style="padding: 2px;">[3,4]</td> </tr> </table>	(4,1)	(2,3)	[3,4]	[3,4]	<u>(3,4)<sup>c*</sup></u>	(1,2) <sup>r</sup>	[3,4]	[3,4]	<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">(4,1)</td> <td style="padding: 2px;">(2,2)</td> </tr> <tr> <td style="padding: 2px;">[3,4]</td> <td style="padding: 2px;">[3,4]</td> </tr> <tr> <td style="padding: 2px;"><u>(3,4)<sup>c*</sup></u></td> <td style="padding: 2px;">(1,3)<sup>r</sup></td> </tr> <tr> <td style="padding: 2px;">[3,4]</td> <td style="padding: 2px;">[3,4]</td> </tr> </table>	(4,1)	(2,2)	[3,4]	[3,4]	<u>(3,4)<sup>c*</sup></u>	(1,3) <sup>r</sup>	[3,4]	[3,4]	<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;"><u>(4,2)<sup>r</sup></u></td> <td style="padding: 2px;">(2,3)</td> </tr> <tr> <td style="padding: 2px;">[4,2]</td> <td style="padding: 2px;">[3,4]</td> </tr> <tr> <td style="padding: 2px;"><u>(3,4)<sup>c</sup></u></td> <td style="padding: 2px;">(1,1)</td> </tr> <tr> <td style="padding: 2px;">[3,4]</td> <td style="padding: 2px;">[3,4]</td> </tr> </table>	<u>(4,2)<sup>r</sup></u>	(2,3)	[4,2]	[3,4]	<u>(3,4)<sup>c</sup></u>	(1,1)	[3,4]	[3,4]
(4,1)	(2,3)																									
[3,4]	[3,4]																									
<u>(3,4)<sup>c*</sup></u>	(1,2) <sup>r</sup>																									
[3,4]	[3,4]																									
(4,1)	(2,2)																									
[3,4]	[3,4]																									
<u>(3,4)<sup>c*</sup></u>	(1,3) <sup>r</sup>																									
[3,4]	[3,4]																									
<u>(4,2)<sup>r</sup></u>	(2,3)																									
[4,2]	[3,4]																									
<u>(3,4)<sup>c</sup></u>	(1,1)																									
[3,4]	[3,4]																									

*Case II:  $(d_1, d_2)$  not Pareto-superior to Nash equilibrium of  $(b_1, b_2)$*

49 (44)	50 (45)	56 (56)																								
<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">(4,1)</td> <td style="padding: 2px;"><u>(2,4)</u></td> </tr> <tr> <td style="padding: 2px;">[2,4]</td> <td style="padding: 2px;">[3,3]</td> </tr> <tr> <td style="padding: 2px;"><u>(3,3)<sup>c*</sup></u></td> <td style="padding: 2px;">(1,2)<sup>r</sup></td> </tr> <tr> <td style="padding: 2px;">[2,4]</td> <td style="padding: 2px;">[2,4]/[3,3]</td> </tr> </table>	(4,1)	<u>(2,4)</u>	[2,4]	[3,3]	<u>(3,3)<sup>c*</sup></u>	(1,2) <sup>r</sup>	[2,4]	[2,4]/[3,3]	<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;"><u>(4,3)<sup>r*</sup></u></td> <td style="padding: 2px;"><u>(2,4)</u></td> </tr> <tr> <td style="padding: 2px;">[4,3]</td> <td style="padding: 2px;">[4,3]</td> </tr> <tr> <td style="padding: 2px;">(3,2)<sup>c</sup></td> <td style="padding: 2px;">(1,1)</td> </tr> <tr> <td style="padding: 2px;">[4,3]/[2,4]</td> <td style="padding: 2px;">[2,4]/[4,3]</td> </tr> </table>	<u>(4,3)<sup>r*</sup></u>	<u>(2,4)</u>	[4,3]	[4,3]	(3,2) <sup>c</sup>	(1,1)	[4,3]/[2,4]	[2,4]/[4,3]	<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;"><u>(4,2)<sup>r</sup></u></td> <td style="padding: 2px;"><u>(2,4)</u></td> </tr> <tr> <td style="padding: 2px;">[4,2]</td> <td style="padding: 2px;">[3,3]</td> </tr> <tr> <td style="padding: 2px;"><u>(3,3)<sup>c</sup></u></td> <td style="padding: 2px;">(1,1)</td> </tr> <tr> <td style="padding: 2px;">[2,4]</td> <td style="padding: 2px;">[2,4]/[3,3]</td> </tr> </table>	<u>(4,2)<sup>r</sup></u>	<u>(2,4)</u>	[4,2]	[3,3]	<u>(3,3)<sup>c</sup></u>	(1,1)	[2,4]	[2,4]/[3,3]
(4,1)	<u>(2,4)</u>																									
[2,4]	[3,3]																									
<u>(3,3)<sup>c*</sup></u>	(1,2) <sup>r</sup>																									
[2,4]	[2,4]/[3,3]																									
<u>(4,3)<sup>r*</sup></u>	<u>(2,4)</u>																									
[4,3]	[4,3]																									
(3,2) <sup>c</sup>	(1,1)																									
[4,3]/[2,4]	[2,4]/[4,3]																									
<u>(4,2)<sup>r</sup></u>	<u>(2,4)</u>																									
[4,2]	[3,3]																									
<u>(3,3)<sup>c</sup></u>	(1,1)																									
[2,4]	[2,4]/[3,3]																									

*Key:*  $(x_1, x_2)$  = (payoff to R, payoff to C) in original game

$[x_1, x_2]$  = [payoff to R, payoff to C] in anticipation game

4 = best; 3 = next best; 2 = next worst; 1 = worst

c = best state C can induce with moving power

r = best state R can induce with moving power

\* = state induced by moving power that is better for both players

Nash equilibria underscored in original and anticipation games

Nonmyopic equilibria (NMEs) circled in original game