2.5 Resolving the paradox III: the metagame solution

Neither deterministic threats nor threats-that-leave-something-to-chance provide a satisfactory resolution of the paradox of mutual deterrence. To resolve the paradox by permitting players to commit to a retaliation strategy is necessarily to jettison the assumption of international anarchy, a core assumption of political realism. And the resolution that relies on a Schellingesque threat to abrogate control solves the paradox, in part, by assuming away the source of the contradiction.

In this section, we explore a third possible way to reconcile the instability of the status quo outcome in Chicken with the observed stability of the superpower relationship during the Cold War period. Based on an idea first suggested by von Neumann and Morgenstern (1944: 100–106), but more fully developed by Howard (1971), this resolution involves an alteration of the underlying game to take into account the possibility that the players might be able to anticipate each other's strategy choice. Presuming that each player bases its own strategy choice on the strategy it expects the other to select, a new game – Howard calls it a *metagame* – is rendered and played "in the heads" of the players prior to the play of the actual game. In the metagame, players choose *metastrategies* rather than strategies. A metastrategy can be thought of as a strategy for selecting a strategy. Stable outcomes of the metagame are termed *metaequilibria*.

To illustrate these concepts, consider once again the game of Chicken, but assume now that State B is able to predict – or thinks it can predict – State A's strategy choice. With this assumption, which is logically equivalent to the assumption that State B selects its strategy after learning A's choice, B's range of choices expands. Rather than having just two strategies (i.e., *C* or *D*), B now has $2 \times 2 = 4$ metastrategies:

1. *C*/*C*: choose *C* regardless of A's choice (*C Regardless*)
2. *D*/*D*: choose *D* regardless of A's choice (*D Regardless*)
3. *C*/*D*: choose *C* if A chooses *C*, *D* if A chooses *D* (*Tit-for-Tat*)
4. *D*/*C*: choose *D* if A chooses *C*, *C* if A chooses *D* (*Tat-for-Tit*),

which gives rise to the *first-level* metagame shown in figure 2.4.

Figure 2.4 can be interpreted in one of two ways: as the game that would be played if B were able to anticipate A's strategy choice (i.e.,

		State B			
		C/C *C Regardless*	*D/D* *D Regardless*	*C/D* *Tit-for-Tat*	*D/C* *Tat-for-Tit*
State A	*C*	(3,3)	(2,4)*	(3,3)	(2,4)
	D	(4,2)*	(1,1)	(1,1)	(4,2)*

(x,y) = payoff to A, payoff to B
4 = best; 3 = next-best; 2 = next-worst; 1 = worst
* = metaequilibria (Nash equilibria)

Fig. 2.4. A first-level metagame of Chicken.

the metagame), or as a sequential (extensive-form) game in which A selects its strategy first.

Notice that there are three metaequilibria in this first-level metagame. Two correspond to equilibria in the original (simultaneous choice) game while the third – (*D*, *D*/*C*) – is strictly a product of the metagame structure. But this new metaequilibrium has a special property that distinguishes it from the other two and, therefore, gives it a singular status: it is the product of B's weakly dominant metastrategy (i.e., *D*/*C*) and A's best response to B's dominant metastrategy (i.e., *D*).[32] Should this equilibrium come into play – and there are good reasons to expect that it, rather than any other, would – A would get its best outcome, and B would get its next-worst outcome.

This preliminary result is interesting for two reasons. First, it shows that the ability to forecast an opponent's strategy does not always help. In Chicken, it actually hurts. And second, it formalizes the view of many decision-theoretic deterrence theorists that the player who seizes the initiative in Chicken wins. Recall that it is on the basis of this observation that decision-theoretic deterrence theorists counsel commitment and related manipulative bargaining tactics.

Metagames, however, do not stop here. Howard now proposes not only that B can anticipate A's strategy choice, *but that A bases its choice*

[32] Recall that a weakly dominant strategy (or metastrategy) is at least as good as, and sometimes better than, any other. For a detailed definition, see footnote 19.

on B's predictions of A's choice. If A conditions its strategy choice on B's metastrategy, it can choose either *C* or *D* for each of B's four metastrategies, which gives State A $2 \times 2 \times 2 \times 2 = 16$ *second-level* metestrategies. For instance, the second-level metastrategy *D/D/C/D* requires A to

1. Choose *D* if B chooses *C/C* (*C Regardless*)
2. Choose *D* if B chooses *D/D* (*D Regardless*)
3. Choose *C* if B chooses *C/D* (*Tit-for-Tat*)
4. Choose *D* if B chooses *D/C* (*Tat-for-Tit*).

A's 16 second-level metastrategies and B's 4 first-level metastrategies imply a $16 \times 4 = 64$ outcome strategic-form game. An abbreviated version of this matrix, listing only non-repetitive metaequilibria, is given in figure 2.5. Notice the increased number of metaequilibria. Among them is one that corresponds to the *Status Quo* outcome CC with payoffs (3,3). This is a significant result because it suggests that if Howard's assumptions are satisfied, the status quo could survive and deterrence could succeed.

The operative word here is "could." There are other possibilities. Still, the metastrategies associated with the (3,3) metaequilibrium of figure 2.5 (i.e., the outcome CC) are explicitly suggestive of the conditions under which deterrence success might occur. Specifically, B's *C/D* metastrategy is a variant of tit-for-tat: cooperate if A cooperates, defect if A defects. So is A's *D/D/C/D* metastrategy. It implies cooperation, but only in response to B's conditionally cooperative tit-for-tat strategy. All of which indicates that mutual cooperation is possible, but only when each player is prepared to cooperate conditionally, that is, when each intends to cooperate should the other player cooperate and – equally important – when each intends to defect should the other defect.

Observe that the metastrategies associated with mutual cooperation (i.e., with stable deterrence) are risky: each carries with it the possibility of a player's worst outcome, DD. But as Howard (1971: 184) argues, if the players are unwilling to run this risk, a compromise equilibrium is not possible. Brams (1975: 44) concurs, adding that the metagame analysis suggests "that a policy of deterrence, by which each side promises retaliation for any untoward acts by the other, is not only desirable from the viewpoint of the players, but stable as well."

If this conclusion holds, the paradox of mutual deterrence is solved. Whether it holds, however, depends on the interpretation given to the

		State B			
		C/C *C Regardless*	*D/D* *D Regardless*	*C/D* *Tit-for-Tat*	*D/C* *Tat-for-Tit*
State A	*C/C/C/C*	(3,3)	(2,4)*	(3,3)	(2,4)
	...	...	...	...	...
	D/C/C/D	(4,2)	*(2,4)**	(3,3)	(4,2)
	...	...	...	...	...
	D/D/C/D	(4,2)	(1,1)	(3,3)*	(4,2)
	...	...	...	...	...
	D/D/D/D	(4,2)*	(1,1)	(1,1)	(4,2)*

(x,y)	= payoff to A, payoff to B
4	= best; 3 = next-best; 2 = next-worst; 1 = worst
*	= metaequilibria (Nash equilibria)
...	= unlisted metastrategies/outcomes

Fig. 2.5. A second-level metagame of Chicken (part).

metaequilibria. Howard's construction is strictly *descriptive*: metaequilibria are established as theoretical possibilities only, and the metastrategies are theoretical statements about the content of the communication necessary to lead to some outcome. In Howard's view, no particular metaequilibrium has special status. Each, therefore, describes a logical possibility in a game between rational players. Which metaequilibrium eventually comes into play depends on what the players expect from one another, or what they communicate to each other in pre-play bargaining and discussion. In the present example, then, mutual cooperation is possible, provided the players are both prepared to cooperate conditionally. But there are also other rational possibilities. For example, should B expect A to select metastrategy *D/C/C/D* and should A expect B to choose metastrategy *D/D* (*D Regardless*), the metaequilibrium italicized in figure 2.5, CD, with payoff (2,4), will occur. This metaequilibrium is best for B and next-worst for A.

Notice that $D/C/C/D$ – or what Howard refers to as the "sure-thing" metastrategy – is weakly dominant for A, giving B good reason to suspect that A will choose it; and since *D Regardless* is B's best response to $D/C/C/D$, A has a good reason to suspect that B will choose it. All this suggests that the metaequilibrium associated with these two metastrategies might well evolve in a game between rational players.

Howard, however, rejects this outcome as *the* solution to the metagame, and denies that any particular reason exists for singling it out. In fact, he argues it would be *foolish* for A to select its sure-thing metastrategy because it induces a worse outcome for A than its "retaliatory" metastrategy $D/D/C/D$. Or in Howard's (1974a: 730) own words, the sure-thing metastrategy is "the strategy of a 'sucker' who invites, and is ready to yield before, the most extreme ultimatum in the possession of his opponent, and is thus willing to surrender his position before any bargaining begins."

But Harsanyi (1974b), hewing to a normative interpretation and insisting on the perfectness criterion, argues that the use of any dominated metastrategy is irrational and, hence, *incredible*.[33] Since a player with a dominant metastrategy always maximizes its expected utility by choosing it, there is no good reason for an opponent to believe that any other metastrategy would be chosen. This, in turn, implies that a player with a dominant metastrategy should choose it.[34] To do otherwise would be to invite calamity.[35] Specifically, if A were to select its retaliatory metastrategy ($D/D/C/D$) and B, anticipating A's sure-thing metastrategy ($D/C/C/D$), selects *D Regardless*, each player's worst outcome, DD, results.

Harsanyi's admonition not to abandon the use of a weakly dominant strategy, especially in a one-shot game, is difficult to ignore: dominant strategies are unconditionally best. But, then, what are we

[33] In chapter 3, we discuss in detail the connection between subgame-perfect equilibria and credibility.

[34] For the particulars of the debate, see Harsanyi (1973, 1974a, 1974b) and Howard (1973, 1974a, 1974b).

[35] It is worth pointing out that the lively three-way debate among Howard, Anatol Rapoport, and Richard Harris over whether the theory of metagames resolves Prisoners' Dilemma also turned on the proper interpretation of Howard's theory. Rapoport's (1967) argument that it does rests on a normative reading was similar to the one advanced by Harsanyi. Howard, refusing to claim anything but a descriptive status for his theory, rejected Rapoport's suggestion. For the full set of citations, see Brams (1975: 39).

to make of Howard's (1974b: 1693) observation that, in Chicken, "what is the *best* strategy from the viewpoint of rationality is the *worst* strategy from the viewpoint of inducement?" In our view, it is simply another way of stating the paradox of mutual deterrence: when conflict is a mutually worst outcome, deterrence stability can only be generated by assuming both "irrational behavior and irrational expectations by the players about each other's behavior" (Harsanyi, 1977: 332).

Rather than solving the paradox of mutual deterrence, then, Howard's methodology highlights it by reformulating it in a way that deepens our understanding. As Brams (1975: 44) adds, "metagame theory specifies precisely, if indirectly, the *content* of the communications and the *nature* of the bargaining necessary to reach compromise." This, of course, is no mean feat. But when interpreted normatively, the theory reveals that compromise, while potentially stable, has no rational basis. In fact, a normative interpretation suggests that B should win, given that B's best response (D/D) to A's dominant metastrategy ($D/C/C/D$) leads to a metaequilibrium at CD. This should be no surprise, since the assignment of a higher-order metastrategy to A is in some sense equivalent to the assumption that State A chooses its strategy with knowledge of B's choice. The observation that the player choosing second in Chicken will lose continues to be robust.

But what if one accepts Howard's strictly descriptive interpretation of metagame theory? In our opinion, the paradox remains unresolved. Without a normative foundation, one is left without an explanation of why, or when, players would transmit the statements necessary to induce and support mutual cooperation.[36]

2.6 Coda

In the previous sections we have examined three proposed resolutions of the paradox of mutual deterrence. Two in fact do resolve the logical

[36] The compromise outcome can be supported, however, if one accepts the "stability by simultaneity" criterion advanced by Fraser and Hipel (1984) in their refinement of Howard's *analysis of options* technique. An otherwise unstable outcome is rendered stable by simultaneity if both players do worse when they switch strategies at the same time. We believe the possibility of simultaneous strategy switches to be so remote that the resolution suggested by this rationality postulate is not germane to our discussion. Thus the puzzle of how to establish the stability of the status quo in Chicken-like contests is not resolved in Fraser and Hipel's system, nor in the more inclusive *graph model* of Fang, Hipel, and Kilgour (1993).