34 Nash Equilibria Are Not Self-Enforcing

1. Introduction

The intuitive basis for Nash's (1951) concept of strategic equilibrium in non-cooperative games has recently received considerable attention. A rationale that has been suggested is that Nash equilibria represent "self-enforcing agreements"; that a pre-play agreement to play a certain strategy tuple will be kept if and only if it is a Nash equilibrium. Several years ago we came across an example that throws doubt on this contention. The example has been cited in various contexts (e.g., Harsanyi and Selten, 1988; Farrell, 1988), but has not heretofore been discussed on its own merits.

Section 2 contains the main example, with an informal discussion. Sections 3 and 4 discuss two other examples, and how they differ from the main one. Section 5 contains a more carefully formulated — though still verbal — argument. Some scenarios for the main example, including one taken from the international relations literature, will be discussed in Section 6. Section 7 summarizes the main points.

2. Example

The game of Figure 1 has two pure Nash equilibria, (c, c) and (d, d). In the absence of pre-play communication, each one has something going for

Research partially supported by NSF grant IRI-8814953 at the Institute for Mathematical Studies in the Social Sciences (Economics), Stanford University, Stanford, CA, USA.

This chapter originally appeared in *Economic Decision Making: Games, Econometrics, and Optimisation: Essays in Honor of Jacques Dreze*, edited by J. J. Gabszewicz, J.-F. Richard, and L. Wolsey, pp. 201–206, Elsevier Science Publishers, Amsterdam, 1990. Reprinted with permission.

	c	d
c	9, 9	0, 8
d	8, 0	7, 7

Figure 1.

it; (c, c) is Pareto-dominant, but (d, d) is much safer.[1] Indeed, since the players cannot communicate, the row player (Alice) may well be uncertain that the column player (Bob) will play c; she might therefore wish to play d, which *assures* her 7, whereas with c she may get nothing. Moreover, if she takes into account that Bob may reason in the same way, she is all the more likely to play d; this makes it still more likely that Bob, too, will play d, and so on. We do not, however, assert that reasonable players *must* play d; only that they *may* do so, that d is not unreasonable or foolish. And for the time being, we assert this only when there is no pre-play communication.

Let us now change the scenario by permitting pre-play communication. On the face of it, it seems that the players can then "agree" to play (c, c); though the agreement is not enforceable, it removes each player's doubt about the other one playing c.

But does it indeed remove this doubt? Suppose that Alice is a careful, prudent person, and in the absence of an agreement, would play d. Suppose now that the players agree on (c, c), and each retires to his "corner" in order actually to make a choice. Alice is about to choose c, when she says to herself: "Wait; I have a few minutes; let me think this over. Suppose that Bob doesn't trust me, and so will play d in spite of our agreement. Then he would still want *me* to play c, because that way he will get 8 rather than 7. And of course, also if he does play c, it is better for him that I play c. Thus he wants me to play c no matter what. So he wants the agreement to play (c, c) in any case; it doesn't bind him, and might increase the chances of my playing c. That doesn't imply that he will necessarily play d, but he may; since he wants the agreement no matter what he plays, the agreement conveys no information about his play. In fact, he may well have signed it without giving any thought as to how actually to play. Since he can reason in the same way about me, neither one of us gets any information from the agreement; it is as if there were no agreement. So I will choose now what I would have chosen without an agreement, namely d."

[1] Technically, (d, d) is *risk dominant* (Harsanyi and Selten, 1988).

	b	f
b	2, 1	0, 0
f	0, 0	1, 2

Figure 2.

Of course, it may be that Alice is not careful and prudent, but impulsive and optimistic,[2] and likes to think that Bob also is. She may then choose c even without an agreement; and so, also with one. We are not saying that rational players will never play c, but only that *agreeing* to do so won't lead them to *do* it. A player might play either c or d, whether or not he has agreed to (c, c); the agreement has no effect, one way or the other. In such circumstances, the agreement should not be called self-enforcing.

3. The battle of the sexes

The above reasoning is not universal; an agreement to play an equilibrium often *is* self-enforcing. Consider, for example, the familiar "battle of the sexes" (Figure 2). Without pre-play communication, the players will be hard put to choose between b (ballet) and f (fight). But if they agree, say, to (b, b), then here they *are* motivated to keep the agreement.

To explain why, consider again how Alice might reason. It is not that she takes the agreement as a direct signal that Bob will keep it. Rather, like in the previous section, she realizes that by signing the agreement, Bob is signalling that he wants *her* to keep it. But unlike in the previous section, here the fact that he wants her to keep it implies that he intends to keep it himself.[3] So for her, too, it is worthwhile to keep it. Similarly for him. *This* agreement *is* self-enforcing.

4. Another example

After reading an early draft of this paper, Professor David Kreps asked if in the game of Figure 3, which is ordinally equivalent to that of Figure 1,

[2] Though still rational!

[3] If Bob plays b, then he would prefer her to play b; if he plays f, he would prefer her to play f. In the previous section, he would prefer her to play c no matter what he does.

	c	d
c	100, 100	0, 8
d	8, 0	7, 7

Figure 3.

(c, c) should not be considered a self-enforcing agreement. The question had us stumped for a while. But actually, the answer is straightforward: Indeed, (c, c) is not self-enforcing, even here. It does look better than in Figure 1; not because an agreement to play it is self-enforcing, but because it will almost surely be played even without an agreement.[4] An agreement to play it does not improve its chances further. As before, both players will sign the agreement gladly, whether or not they keep it; it therefore conveys no information.

5. Discussion

To say that a game is non-cooperative means that there is no external mechanism available for the enforcement of agreements. Thus when the time comes to choose an action, the players are assumed to act on the basis of the existing incentives. Therefore an agreement is effective only if it changes the incentives that obtain in the absence of the agreement.

Incentives can be changed by changing either the payoffs or the information of the players. The agreements being discussed here do not change the payoffs; the payoffs for any particular strategy tuple remain the same, whether or not it violates the "agreement". To be effective, therefore, an agreement must change the players' information; specifically, their information about how the others will play.

Information about an event E is acquired by observing a parameter that depends on whether or not E obtains. If the parameter does not really depend on E — has the same value whether or not E obtains — then observing it yields no information about E.

In the games of Figures 1 and 3, Alice is interested in knowing what Bob will play; we may take E to be the event "Bob will play c". The parameter she observes is whether or not he "agrees" to (c, c). But this parameter is

[4] Here (c, c) is risk-dominant (cf. Footnote 1).

the same no matter what Bob plays; it is always to his advantage to agree to (c, c). Therefore the agreement yields no information about what he will really play. Since the agreement is important only for the information it yields, and yields no information, it is as if it had not been made.

6. Scenarios for the main example

The game of Figure 1 is sometimes called the "stag hunt".[5] Two men agree to hunt a stag. To succeed, they must go along separate paths, giving the task their undivided attention. On the way, each has the opportunity to abandon the stag hunt and hunt rabbits instead. If he does so the number of rabbits he bags increases if the other continues to hunt the stag. Both would prefer it if both hunted the stag, since it is more valuable than a bag of rabbits. But each fears that each mistrusts the other, that the mistrust breeds more mistrust, and so on.

In the international relations literature, the game has been called the "security dilemma" (Jervis, 1978). Two countries between which there is tension are each considering the development of a new, expensive weapons system. Each is best off if neither has the system, but would be at a serious disadvantage if only the other had it. Can either side afford not to develop the system?

Some closely related games played a role[6] in the controversy about NTU values between Roth (1980, 1986) and Aumann (1985, 1986). There, each of the two players may be offered a deal by an outside party; if both refuse, they can make a better deal with each other, but each fears that the other will close with the outside party before they get a chance to talk with each other.

7. Summary

A non-binding agreement can affect the outcome of a game only if it conveys information about what the players will do. Directly, the information that such an agreement conveys is not that the players will keep it (since it is not binding), but that each wants the other to keep it. In the battle of the sexes (Figure 2), an agreement to play (b, b), say, conveys the information

[5] We have not succeeded in hunting this story down to its source.

[6] See Aumann (1985, Sections 5 and 6, pp. 670–673).

that each player prefers the other to play b; this implies that each will play b himself, and so the agreement is self-enforcing. But in the games of Figures 1 and 3, each player *always* prefers the other to play c, no matter what he himself plays. Therefore an agreement to play (c, c) conveys no information about what the players will do, and cannot be considered self-enforcing.

References

Aumann, R.J. (1985) 'On the non-transferable utility value: A comment on the Roth-Shafer examples', *Econometrica*, 53:667–677 [Chapter 61c].

Aumann, R.J. (1986) 'Rejoinder', *Econometrica*, 54:985–989 [Chapter 61e].

Farrell, J. (1988) 'Communication, coordination, and Nash equilibrium', *Economics Letters*, 27:209–214.

Harsanyi, J.C. and R. Selten (1988) *A General Theory of Equilibrium Selection in Games*. Cambridge and London: MIT Press.

Jervis, R. (1978) 'Cooperation under the security dilemma', *World Politics*, 30:167–214.

Nash, J.F. (1951) 'Non-cooperative games', *Annals of Mathematics*, 54:286–295.

Roth, A. (1980) 'Values for games without side payments: Some difficulties with current concepts', *Econometrica*, 48:457–465 [Chapter 61a].

Roth, A. (1985) 'On the non-transferable utility value: A reply to Aumann', *Econometrica*, 54:981–984 [Chapter 61d].