ALLAN GIBBARD

WEAKLY SELF-RATIFYING STRATEGIES: COMMENTS ON McCLENNEN

(Received 5 August, 1991)

In Edward McClennen's rich paper, I want chiefly to talk about neutral equilibria: situations in which each person's strategy — usually mixed — is a best response to the other's, but not the only best response. In many games (in the game-theoretic sense), a neutral equilibrium with mixed strategies is the only equilibrium there is. This is strange, McClennen and I both think. A mixed strategy is rational, the standard theory tells us. By expected utility reasoning, then, any of the moves this mixed strategy gives you any chance of making is rational. Moreover, any other lottery among these moves is rational. Yet the theory picks out a unique mixed strategy, and tells you to adopt it.

In these cases, your equilibrium strategy is the only strategy that is *self-ratifying*, in the sense that if you know that it is the strategy you are adopting, then you will rationally think that it is the strategy that holds out best prospects. But it is only weakly self-ratifying: If you adopt it, you won't think it is the unique strategy that holds out best prospects. It recommends itself, but it recommends alternative strategies too. Adopting your equilibrium strategy, you will think that these alternative strategies hold out equally good prospects. But these alternative strategies are not self-ratifying: for any of them, if you had known that it was the strategy you were adopting, then you would not have thought it held out best prospects. Only one strategy recommends itself, but it recommends other strategies equally. The strategy is uniquely but weakly self-ratifying.[1]

In the case of two person zero sum games, games were one player's gain is another's equal loss, McClennen thinks the standard game theorists' solution is correct for all its strangeness, but that it cannot be supported by expected utility reasoning, the kind of reasoning that applies to games against nature. Now I keep agreeing that there is something strange about neutral equilibria, and I'll be saying more

Philosophical Studies **65**: 217—225, 1992.

about this strangeness at the end of my remarks, without resolving the matter.

McClennen, though, draws a grand moral that I don't think is justified. He thinks there is a fundamental difference between games against nature and games against other rational agents. One, he says, calls for "parametric rationality" — taking the parameters of one's situation as given. The other calls for strategic rationality. I want to say this is wrong. What gives rise to uniquely but weakly self-ratifying strategies is not, necessarily, that one is playing against a rational agent. It is that which strategy one adopts is evidentially significant. In the case of a game, what one does serves as evidence for what one's opponent is expecting one to do. But this is not the only way one's strategy can indicate one's environment.

The story I will tell to illustrate this point is far-fetched, but my point is a theoretical one, and so a far-fetched story will do. Imagine this: Smoking, it turns out, does cause cancer, but only in people with a certain gene. This gene, indeed, has two effects: First, as I said, it makes one susceptible to smoking's causing cancer. Second, though, it makes one reckless in situations precisely like this, so that in this situation, if one enjoys smoking, one will smoke despite the danger. Knowing all this, the question is, what is it rational to do if you somewhat enjoy smoking, and very much don't want to get cancer?[2] If you recklessly smoke, that indicates you have the gene, so that with you, smoking causes cancer. If you cautiously refrain from smoking, that indicates you don't have the gene, so that you could enjoy smoking and still not get cancer.

You might adopt a mixed strategy, and give yourself some intermediate chance of smoking. For this case, we have to elaborate the story. Imagine in addition, that the more reckless you are — the higher a chance you give yourself of smoking — the more likely you are to have the gene. Suppose the probability of having the gene is a continuous function of how recklessly you gamble. Then there will be a probability of smoking such that, if you know that it is the probability you are adopting, the prospects you rationally envisage are equal whether you smoke or not. This strategy is self-ratifying, but only weakly. It recommends itself, but equally recommends any other probability mixture of smoking or not smoking.

The big divide, then, is not between games against nature and games against rational agents. The peculiarities come when one's choice strategy gives evidence about features of the environment that one's actions cannot affect. Then strange things can arise, including the existence of a uniquely but weakly self-ratifying strategy.

1. TWO KINDS OF CONDITIONALS

McClennen has another argument against standard treatments of game theory, an argument that does not rely on the strange features of uniquely but weakly self-ratifying strategies. He discredits an assumption required for the standard arguments in game theory, by arguing that this assumption would allow an equally good argument for a choice rule that is ridiculous: to evaluate each strategy by asking how good things would be if that were the strategy your opponent correctly foresaw you adopting. Act, in other words, as if you were to choose your strategy first, and then your opponent, observing your strategy, would respond. McClennen and I agree that this way of thinking is defective. It has you acting as if you were choosing first, when in fact you aren't. In a game of chicken, if both parties think this way, they die in a head-on crash.

McClennen, though, thinks the manifest folly of this way of thinking discredits the standard assumptions of game theory. The assumption he attacks is this: that you the rational player are convinced, throughout your reasoning, that when you finally conclude what it is rational to do, your opponent will have foreseen your conclusion, and already expect you to do it. You expect that the other player has "psyched you out." Call this the *Mutual Predictability Assumption.* McClennen claims that from this assumption would follow the ridiculous decision rule of acting, in a simultaneous game, as if you are moving first — and so the assumption must be untenable.

I shall argue that Mutual Predictability has no such consequence. To see this, we must understand exactly what Mutual Predictability says, and it will help if we review its rationale. An early point in your reasoning, you think there is a fully convincing theory of how to play the game rationally, and that you will follow the theory and your opponent will expect you to. At this point in your reasoning, though,

you don't know what the theory is. You thus think you will choose using the right theory, once you have figured out what that theory is, and that your opponent too will figure out the right theory and anticipate your using it. You think, therefore, that when you have decided what to do, your opponent will already be expecting it, and respond rationally to the move he predicts you will be making.

The rationale supports this thought, but not another with which it might be confused. Here is something you cannot reasonably think: *That no matter what you might do, if you were to do it, your opponent would anticipate it.* The confusion stems from two readings of the conditional, "If I adopt strategy A, then A is the strategy my opponent will expect me to adopt." First, there is the epistemic reading, on which it may be a perfectly reasonable thing to accept. "If I'm going to do A, then he expects me to do A." Fully convinced that I will do whatever is rational, and that he knows what is rational and expects me to do it, and unsure whether A or something else is the rational thing to do, I am set, on learning I will do A, to think he expects me to do A. The unacceptable reading is the subjunctive one: "If I were to do A, then A would be what he expects me to do." I don't know whether A is the rational thing to do. If something else — B, let us say — is the rational thing to do, he anticipates my doing B. Thus if I were to do A, then he would get me wrong.

Now Robert Stalnaker discovered that subjunctive conditionals have a special role in rational decision. The point has been controversial, but William Harper and I are among the people who think this is a genuine discovery — and even fairly obvious, once one thinks about it enough. It counts against an act that if I were to do it, something bad would ensue, and in favor of the act that if I were to do it, something good would ensue.[3]

Let me illustrate the distinction with a parallel decision problem, where again a conditional can be given either of these two kinds of readings, and confusing the two could lead to folly. I am good at arithmetic: I think slowly, but when I come to an answer, the answer is right. I have won two prizes, worth \$58 and \$67. I can collect my prize only if I answer correctly what I am owed. I am to answer on a multiple choice test; possible answers \$110, \$115, \$120, \$125, \$130. Now knowing I am good at arithmetic, I think, epistemically, that if I'm going

to conclude the sum is $130, then that is the right answer, and $130 is what I will get. Likewise, if I'm going to conclude that the sum is $125, then $125 is what I will get. These epistemic conditionals are fine. What I cannot reasonably do is go on to reason like this: I prefer $130 to $125, and since if I'm going to mark $130 on the test, then that is what the sum is, it follows that I do better to mark $130. To reason this way is to endorse wishful thinking. The epistemic conditional is the wrong one to use. Rather, I must guide myself by subjunctive conditionals. But I don't think, subjunctively, that if I were to mark $130 on the test, the sum would really be $130, so that $130 is what I would get. For all I know so far, the sum isn't $130, and if it isn't, then if I were to conclude that the sum is $130, I would get nothing.

Things go likewise in the game-theoretic case. I can reasonably accept an epistemic conditional. I'm convinced that I'm rational, and that he knows it, and that he too will have figured out what is rational. So I accept this: "If I am going to choose A then he expects me to choose A." I can't use this, though, as a reason for choosing whatever I'd most like him to correctly believe I was choosing. That would endorse driving straight on in a game of chicken. What matters directly for decision is whether I accept this subjunctive conditional: "If I were to drive straight on for sure, he would expect me to do so — and so would swerve." If I were reasonably convinced of this, I could reasonably drive straight on for sure, knowing he would swerve. But I can't reasonably be convinced of this — for I haven't yet concluded whether driving straight on is the rational thing to do. At that point, I know only that *if* it is the rational thing to do, then he expects me to do it. Thus if it is what I am going to do, then it is what he expects me to do. But it doesn't follow that *were* I to do it, that is what he would have expected me to do — that would be a silly thing to accept.

Now to my eye, McClennen's "Maxilor Condition", as he states it, gets an epistemic reading. It reads to me as a plausible condition, and one that could not be used to support acting, in a simultaneous game, as if one were moving first. Its upshot is that one's strategy must be a best response to one's opponent's best response. It does not say that one's strategy must be best on the false assumption that no matter what strategy one were to choose, one's opponent would give a best response to it. Yet this last is what constitutes acting as if one were moving first,

and one's opponent were going to observe one's move and give a best response. This last is the rule that tells both drivers in a game of chicken to drive straight on — and so to crash.

The standard requirement on a rational strategy X is this:

> X must do at least as well against the best reply to X as any alternative Y does against the best response to X.

This is different from the silly requirement that produces a head-on crash in the game of chicken:

> X must do at least as well against the best reply to X as any alternative Y does against the best response to Y.

The silly requirement does not follow from the standard requirement. I read the "Maxilor Condition", as McClennen states it, as yielding the standard requirement, not the silly one.

Now the standard argument in game theory uses the right kind of conditionals in the right way. It uses an epistemic conditional, and it uses it correctly: to drive an indirect argument, not to say what would happen if one acted one way or another. How do we rule out a pure strategy? I say, "If A is rational for me, then he knows that A is rational for me, and he is going to give his best response to that. But A is not my best response to this, and so A is not rational." This uses only well founded epistemic conditionals. It nowhere assumes the fallacious subjunctive claim, that if I were to do A, he would give his best response to A. But it is this subjunctive kind of conditional that is needed to drive the crazy maxilor argument.

2. STRANGE EQUILIBRIA

What do I want to say, then, about the puzzle of a uniquely but weakly self-ratifying equilibrium? I'm not at all sure what to say, especially in the smoking case. There, the unsatisfactory way of reasoning I have been sketching seems forced on us. The only way out I can see is this: Could it be that there is no ideally rational thing to do in the circumstance? However rational one is, we can argue, one needs to be more rational. A more rational being, after all, could look at you, conclude

what you are disposed to do, deduce from that what your genetic makeup is, and act accordingly. You yourself are not guileful enough to do so.

In Ghana, girls play a game called *ampeh*. Clapping their hands twice, each girl jumps first on two feet and then on one foot. One girl tries to match the foot the other girl jumps on; the other tries to avoid being matched. Now imagine two robots programmed to play ampeh — though doubtless, without the grace and exuberance of young girls at play. The robots are transparent, in the sense that each can "see" the other's program. Then we know that at least one program is incapable of predicting which foot the other will choose — or at least, to do so long enough before the second clap to respond optimally. For call the programs Amma and Akua, and suppose Amma is trying for a match and Akua for a non-match. If Amma could predict Akua in time, she would match. If Akua could predict Amma in time, she would mismatch. So it can't be the case both that Amma can predict Akua in time and Akua can predict Amma in time. Yet each has full information about the other: they are transparent. What must be happening is that at least one of them is insufficiently powerful a reasoner to figure out the other on time, even given a full view of the other's workings. A sufficient powerful algorithm could predict them both — and so the moral is, there is no such conceivable thing as an ideally powerful algorithm. For every possible algorithm, there is a more powerful one.[4]

All this is for algorithms that cannot randomize. They can pseudo-randomize: they can perform determinate calculations that pretty well mimic randomization. But in one way, they do not mimic true randomization: a sufficiently powerful observer could predict the outcome of the process.

Now we had thought that with true randomization, we weren't forced to the conclusion that there is no such thing as ideal rationality. Or at least, it seemed we could think that for a given problem, there is such a thing as being rational enough — being a powerful enough calculator — no matter what one's opponent is like. The same thing seems to go for algorithms that use opaque pseudo-randomizers: who are able to keep each other ignorant of the calculations they use for pseudo-randomization. We should find these results striking and surprising. And indeed,

faced with the paradox of a uniquely but weakly self-ratifying strategy, perhaps we should say, these results turned out to be too good to be true.

In the game-theoretic case, though, I'm not sure we are forced to this conclusion — and this for reasons McClennen gives. The crucial assumption is one of *determinateness:* that from one's situation and one's rationality, it follows what one's strategy will be. There seems to be no reason to accept this assumption. Why shouldn't the correct theory of rationality endorse more than one choice as equally rational? In that case, your opponent may know your situation, know that you are rational, and know what the correct theory of rationality says about our situation — and still not know what you will do. That is to say, he may not only not know what you will end up doing, but — even what probabilities you will give yourself of doing various different things.

How, then, would we get a mixed strategy equilibrium? It might not be that any one player adopts the mixed strategy that game theory recommends. It might even be that each adopts a pure strategy, but that neither player knows which pure strategy the other is adopting. Let Dick and Jane engage in a zero sum game with mixed strategies in equilibrium. What matters for this equilibrium is that Dick have the right subjective probabilities for what Jane will do, not that these be the actual probabilities with which she will do these things. The same applies, of course, to Jane psyching out Dick. Dick and Jane may even be incapable of randomizing, and know it. But for Dick to have the right subjective probabilities for what Jane will do, all we need is that Dick doesn't know if he is facing a Jane type 1 who will do act A, or a Jane type 2 who will do act B.

I don't know how to work out a theory of this kind so as to be fully satisfactory. But it does have a virtue: It does not issue in the strange requirement that in a situation in which Jane finds acts A and B to offer equal prospects, she must, to be rational, choose some particular probability mixture of the two. It allows that any probability mixture would be equally rational — even A for sure or B for sure.

It might be, then, that game theory is an easier nut to crack, helping ourselves to the standard assumptions of parametric rationality, than is the smoking example I gave — though the smoking example is a game against nature. The big divide in the theory of instrumental rationality is

not between games against rational players and games against nature, but between actions that are diagnostic of one's environment and actions that are not. More precisely, the tricky cases are ones in which one's action will give one evidence about features of one's environment the action cannot affect. Game theory has strange features because its standard assumptions, seemingly well founded, give rise to cases of this sort.

NOTES

[1] Richard Jeffrey discusses the ratifiability of decisions in the Second Edition of *The Logic of Decision* (Chicago: University of Chicago Press, 1983), pp. 15—20.
[2] This style of example and its use owes much to Robert Stalnaker. See his letter to David Lewis of May 21, 1972, printed in William L. Harper, Robert Stalnaker, and Glenn Pierce, *Ifs* (Boston: D. Reidel, 1981), pp. 151—2.
[3] Stalnaker, letter to David Lewis, cited above. Harper and I proselytized Stalnaker's proposal in "Counterfactuals and Two Kinds of Expected Utility", C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory* (Boston: D. Reidel, 1978), 125—163.
[4] I draw this argument from David Lewis and Jane S. Richardson, "Scriven on Human Unpredictability", *Philosophical Studies* 17 (1966), 69—74.

Department of Philosophy
University of Michigan
Ann Arbor, MI 48109
USA