REVIEWS AND COMMENTS

With the intent of stimulating discussion, this section is reserved for book reviews, comments, and letters; your input is welcome. By nature, this material may be subjective, reflecting the opinions of the authors; your responses are therefore encouraged.   © 1998 Academic Press

Counter-Counterfactuals

Itzhak Gilboa*

Boston University, Boston, Massachusetts, 02215; and Tel-Aviv University, Tel-Aviv, Israel

THE PROBLEM

"It will definitely rain this morning. However, if it doesn't, it will definitely be very cold this afternoon." A strange prediction, indeed. If the forecaster believes it will rain, why does she bother to talk about what will happen if it doesn't? And, in case it doesn't rain in the morning, should we take her secondary prediction seriously? After all, if this forecaster used to think that it would rain, and was proven wrong, why should we believe that she would now be any more accurate in her predictions? Wouldn't we be better off relying on another forecaster, one who had said it wouldn't rain in the first place?

The prediction that, if it doesn't rain, it will be cold in the afternoon, may be viewed as a counterfactual, i.e., as a substantive conditional statement whose antecedent is known to be false. But this is a counterfactual of a special kind. It is not yet known that it rained in the morning. Rather, the prediction that it will rain is part of the theory to which the counterfactual belongs. As long as we believe in the theory's primary prediction (namely, that it would rain), the conditional statement is a counterfactual. But this counterfactual lurks in the background, waiting for the theory's primary prediction to fail, and then it would become factual

and assume the title "the theory's prediction." Let us refer to such conditionals as "counter-counterfactuals."[1]

Scientific theories are typically neither required nor allowed to specify what would happen if they were refuted, that is, to use counter-counterfactuals.[2] Yet, game theory seems to be replete with counter-counterfactuals. A noncooperative "solution concept" involves a choice of a strategy for each player, and it thus specifies what would happen, as well as what would happen if what should have happened did not happen, and so forth. It is as if game theorists are never surprised to see their predictions fail; as if we are proven wrong time and again, but we still have the audacity to keep making predictions for the future.

WHY US?

What have we done to deserve this? Why, of all sciences, does game theory have to be the one that deals with counter-counterfactuals? To understand our fate, it may be useful to see why other sciences fare better. The natural sciences do not deal with conscious decision makers. Thus they need not worry about the possibility that the subjects of their theory would decide to render it false. The same is true of much of economics and sociology and, indeed, of parts of game theory that deal with large populations: as long as a single decision maker cannot refute the theory, there is no need to specify what would happen if the theory were wrong.

Psychology also deals, at times, with individual decision makers, but psychological theories need not involve counter-counterfactuals. For instance, the behaviorist stimulus–response paradigm yields predictions of people's behavior without any reference to conscious decision making. The dominant paradigm in game theory, however, differs from behaviorism in that it attempts to model *reasoned* choice. Hence, game theory is bound to describe what the decision maker thinks would occur should she take a

---

[1]Observe that, if it does not rain, we will argue that the theory has been refuted. One cannot resort to a probabilistic interpretation of the theory in order to accommodate the data. Should probabilistic predictions be meaningful, they have to be falsifiable, say, in a statistical sense. And then our discussion should begin when the predictions have failed statistically. For this reason, "trembling hand" arguments are hardly convincing as a resolution of backward induction paradoxes.

[2]In a sense, the scientific method itself does involve counter-counterfactuals. The methodology of science provides some guidelines for replacing failed theories by new ones. Consumers of such theories might wonder why they should believe the same scientists who failed in the past. Indeed, claims of this nature are often heard regarding economics and even medicine. Yet, a particular scientific theory does not typically contain counter-counterfactuals.

certain action, for each and every possible action, and not only for the one that she is predicted to take.[3, 4]

Yet a theory of reasoned choice need not engage in counterfactuals itself. It may specify what players think would happen were the theory refuted, without subscribing to these counterfactual beliefs. But such a theory will not be consistent with the assumption that the players share the theory of the game with the modeler. While game theoretic predictions need not be common knowledge among the players, it is bothersome to think that in principle they cannot.

To sum, the predicament of game theory follows from the fact that its goal is to model reasoned choice of individual decision makers, in a way that is consistent with the implicit assumption that the theory itself is commonly known. As long as game theory remains committed to this goal, it is bound to engage in counter-counterfactuals. Whatever the theory's primary predictions, it has to specify secondary predictions, describing the reasoning that leads the players to play according to the primary predictions. But players may refute the primary predictions, thereby calling into doubt also the secondary predictions that are part of the same theory. It is therefore not quite clear why players should follow the primary predictions, which were supposedly derived from the secondary ones.

SO WHAT SHOULD WE DO?

A prediction "A, but if not A, then B" should be viewed as a multilayered, admittedly self-referential theory that can rise from its ashes and predict the future following its own demise. The interpretation of its second clause should be that, if A is found false, B will become the commonly known prediction of all players involved, *despite the fact that they will have witnessed a violation of the primary predictions*.[5] But, says the reader, here is the problem again: how can they still believe in B, which follows from the theory, now that the theory has been falsified? Well, we should reply, what else did you think that this clause meant in the first place? What could possibly be the meaning of "but if not A, then B" if we can only apply it when A is true?

---

[3]Reasoned choice does not require maximization of expected utility or any other notion of behaviorally-defined "rationality." It only requires that the decision maker consider all options and arrive at a decision by some process of reasoning, to be modeled by the theory.

[4]Applications of game theory to biology are, of course, an exception.

[5]This claim is primarily normative: I argue that this is the interpretation of counter-counterfactuals that we should adopt. But I do not view it as revolutionary. I suspect that this interpretation also conforms to the way many game theorists tend to view strategy profiles.

CK OF RATIONALITY OF AGENTS AND OF PLAYERS

Following this type of reasoning, one may argue that common knowledge (CK) of rationality of agents leads to backward induction. The assumption that each and every agent is rational and that this fact is common knowledge among them can only mean that the backward induction solution will be followed in every subtree, *no matter how many times the backward induction solution has been refuted in the past*. Indeed, what else can the assumptions of rationality of the agents, and of common knowledge thereof, mean unless one applies them for predicting behavior once the relevant agents finally come into existence and have to make choices? What is the point in believing that certain agents are rational unless they are called upon to play?[6]

By contrast, CK of rationality of players may mean more than one thing. For instance, some argue that rationality of a player implies the rationality of all her agents. Others may find rationality meaningful even if it is assumed only at some nodes, say, those that are predicted to be reached, and not necessarily at others. (See Binmore, 1987; Bicchieri, 1988, 1989; Reny, 1992; Ben-Porath, 1992; Samet, 1994; Aumann, 1995; Binmore, 1996; and Aumann, 1996.)

CK of rationality of agents has been criticized for its use of counter-counterfactuals: how can one maintain this assumption once it is patently violated? As argued above, this criticism applies to *any* game theoretic prediction. Why then has CK of rationality drawn such heavy fire? A possible answer is that our belief in a scientific theory is intensional, that is, that it depends on the way the theory is represented. When the theory's predictions at every node of the tree are spelled out, we tend to believe that the game theorist has made the prediction at each node precisely for the case that this node is reached. We then treat these predictions as what the theorist "really" meant, even if they are counter-counterfactuals in the context of the theory as a whole. By contrast, when the theory's predictions follow from a general principle, we are justifiably suspicious that the game theorist has not devoted enough thought to some nodes.

Yet, there is nothing wrong with the theory that all agents are rational and that this fact is CK among them. Should a game theorist insist that she believes in this theory and that she has seriously thought about it for each and every node, we have no choice but to admit that in her model

---

[6] Using arguments of the type "What could 'i knows that j knows that ... that k is rational' mean but ..." one derives the backward induction solution. Any other interpretation of CK of rationality would render at least one statement about agents' knowledge (of a certain order) of rationality (of a certain agent) devoid of empirical content.

backward induction would prevail and that there is nothing peculiar about her model, or rather, nothing that is more peculiar than game theory itself.

BETWEEN AUMANN AND BINMORE

Aumann (1995) proves that CK of rationality of players implies that they would play the backward induction solution. His paper has received much criticism, some of which is in print. Specifically, Binmore (1996) has criticized Aumann's definition of rationality and of CK thereof. Aumann (1996) has shown that there is nothing strange about his definitions. Indeed, he can prove his result with the standard definitions of Bayesian rationality and of common knowledge. Moreover, everyone seems to agree that CK of rationality is an unrealistic assumption and that the argument is not about the verisimilitude of the backward induction result. Finally, no one doubts the mathematical correctness of Aumann's theorem.

So what is the argument about? And why are some readers unhappy about Aumann's result? I believe that the answers to these questions have to do with our coming to terms with counter-counterfactuals. While game theoretical predictions are inherently counter-counterfactual, this fact has not been explicitly acknowledged in the literature. Backward induction paradoxes are probably the simplest examples of the problematic counter-counterfactual nature of all strategy-profile predictions. But the criticism of Aumann's result has little to do with rationality or with backward induction per se.

Aumann (1995) formalizes the notions of CK and of rationality. But his model does not formalize counter-counterfactuals. That is, the self-referential nature of game theoretic predictions is not reflected in the mathematical model Aumann employs. Further, his interpretation of the model does not attempt to say anything about what would happen if the theory were refuted. Indeed, this was not his goal. But this is what some readers would have liked to see: a theory that would explicitly and shamelessly say "should this theory be proven wrong in the morning, it predicts that in the afternoon . . ." Perhaps only a formal model that captures self-referential theories of this nature can settle the backward induction issue and can thereby enable us to set our minds to theories that are not only consistent but also realistic.

REFERENCES

Aumann, R. J. (1995). ''Backward Induction and Common Knowledge of Rationality,'' *Games Econ. Behav.* **8**, 6–19.

Aumann, R. J. (1996). ''Reply to Binmore,'' *Games Econ. Behav.* **17**, 138−146.

Ben-Porath, E. (1992). ''Common Belief in Rationality in Extensive Form Games,'' *Rev. Econ. Stud.*, to appear.

Bicchieri, C. (1988). ''Strategic Behavior and Counterfactuals,'' *Synthese* **76**, 135−169.

Bicchieri, C. (1989). ''Self Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge,'' *Erkenntniss* **30**, 69−85.

Binmore, K. (1987). ''Modeling Rational Players I,'' *Econ. Philos.* **3**, 179−214.

Binmore, K. (1996). ''A Note on Backward Induction,'' *Games Econ. Behav.* **17**, 135−137.

Reny, P. (1992). ''Rationality in Extensive Form Games,'' *J. Econ. Perspectives* **6**, 103−118.

Samet, D. (1994). ''Hypothetical Knowledge in Games of Perfect Information,'' mimeo.