econjournalwatch.org

*Econ Journal Watch,
Volume 2, Number 1,
April 2005, pp 97-104.*

SYMPOSIUM ON INFORMATION AND KNOWLEDGE IN ECONOMICS

Why the Distinction Between Knowledge and Belief Might Matter

KEN BINMORE*

THE ENGLISH LANGUAGE DIMLY RECOGNIZES THAT KNOWING something is not the same as believing it sure to be true. Ptolemy believed it sure that the sun revolved around the earth, but we would not say that he *knew* the sun revolved around the earth. Philosophers commonly claim that the difference lies in the fact that knowledge should be regarded as justified true belief, but such an attempt at a definition has had little influence in rational choice theory, presumably because the question of what should count as a justifying argument is left hanging in the air.

In this brief essay, I want to argue that there are different reasons why in rational choice theory it may sometimes be worthwhile to distinguish between knowledge and belief-with-probability-one. I suggest employing the latter for situations in which the model within which an event is believed to obtain with a subjective probability of one is best regarded as the limit of a sequence of models in which the event in question is believed to obtain with a subjective probability less than one. An example may help to explain my reasons for making this suggestion.

Alice is a perfectly rational decision-maker who values her own safety. She therefore won't step in front of a car when crossing the road. I am so sure of my facts that I attribute probability one to this assertion. But what was my reasoning process in coming to this conclusion? I have to

* Emeritus Professor, University College London.

contemplate Alice comparing the consequences of stepping in front of a car with staying on the kerb. But how can Alice or I evaluate the implications of the former event, which we know is impossible? In mathematical logic, anything whatever can be deduced from a contradiction.

The answer is that the material implication of mathematical logic is the wrong reasoning tool in such situations. We have to appeal instead to what philosophers call a *subjunctive conditional* when faced with the need to evaluate counterfactuals. Since counterfactuals are counter to the facts of our world, they are impossible in our world. Reasoning with them therefore requires postulating another world in which they are possible. For example, Alice will not step in front of a car in this world, but she can imagine another possible world in which a momentary lapse of attention results in her making this mistake. The consequences of her stepping in front of a car in this possible world of mistakes are so bad that she exercises great care not to make any mistake in the actual world, thereby confirming our hypothesis that she does not make that mistake.

INTERPRETING COUNTERFACTUALS

But what possible world should we use when interpreting a subjunctive conditional? The standard reference in philosophy is Lewis's (1976) *Counterfactuals*, but more practical advice is offered by Selten and Leopold (1982). They suggest expanding whatever model we are using to represent the actual world by introducing new factors that make events that are impossible in the old model become merely improbable in the new model. One can then interpret a statement that conditions on a counterfactual event in the original world as the limiting case when the probability of the corresponding event in the expanded world is allowed to become vanishingly small.

In giving similar advice for conditioning on zero-probability events, Kolmogorov (1950) gives examples to show that we can be led to different answers by expanding our model in different ways. For this reason, it is vital to be aware of the context in which counterfactuals are introduced, because there is nowhere but the context to look when deciding what expansion of the basic model is appropriate (Schelling 1960: 53-118). To say that Alice believes something with probability one invites us to recognize that things might have been different in some other possible world—and therefore that the *context* within which we chose our current model of the actual world may well be relevant.

Saying that Alice *knows* something may then be reserved for events that will be kept constant in all the possible worlds invoked in interpreting relevant counterfactuals.

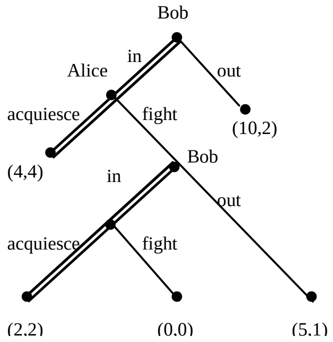
The relevance of these considerations to rational choice theory is highlighted by Aumann's (1995) claim that common knowledge of rationality implies that play in a finite game of perfect information must lie on the path to which one is led by applying the principle of backward induction.

A rational player stays on the equilibrium path in a game because of what *would* happen if he *were* to deviate. We cannot therefore understand what rationality is without interpreting a subjunctive conditional in which the conditioning event is counterfactual. How we interpret the counterfactual depends on the context. For example, if the expanded world we introduce to interpret the counterfactual involves mistakes, we may ask whether the mistakes are `typos' or `thinkos'. If the former, a mistake made early in the game can be dismissed as a transient phenomenon without relevance for the player's future play. If the latter, then mistakes are likely to be correlated, with the result that the arguments offered in favor of backward induction cease to apply (Binmore 1987, Fudenberg and Levine 1993).

CHAIN-STORE PARADOX

The following simplification of Selten's Chain-Store paradox may clarify this point. Alice's chain of stores operates in two towns. If Bob sets up a store in the first town, Alice can acquiesce, or fight a price war. If he later sets up another store in the second town, she can again acquiesce or fight. If Bob chooses to stay out of the first town, we simplify by assuming that he necessarily stays out of the second town. Similarly, if Alice acquiesces in the first town, we assume that Bob necessarily enters the second town, and Alice again acquiesces. The doubled lines in Figure 1(a) show that backward induction leads to the play [*ia*], in which Bob enters and Alice acquiesces. The same result is obtained by successively deleting (weakly) dominated strategies in Figure 1(b).

Figure 1: A simplified Chain Store



(a)

	ii	io	oi	oo
aa	4,4	4,4	10,2	10,2
af	4,4	4,4	10,2	10,2
fa	2,2	5,1	10,2	10,2
ff	0,0	5,1	10,2	10,2

(b)

Suppose Alice reads an authoritative book on game theory which says that the play [*ia*] is rational. Alice will then arrive at her first move with her belief that Bob is rational intact. To check that the book's advice to acquiesce is sound, she needs to predict what Bob would do at his second move in the event that she fights. But the book says that fighting is irrational. Bob would therefore need to interpret a counterfactual at his second move: If a rational Alice behaves irrationally at her first move, what would she do at her second move?

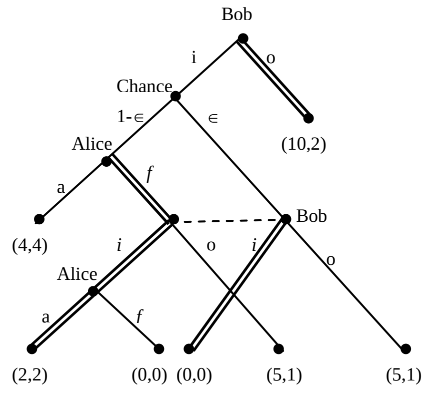
There are two possible answers to this question: At her second move, Alice might acquiesce or she might fight. If she would acquiesce, then it would be optimal for Bob to enter at his second move, and so Alice should acquiesce at her first move. In this case, the book's advice is sound. But if Alice would fight at her second move, then it would be optimal for Bob to stay out at his second move, and so Alice should fight at her first move. In this case, the book's advice is unsound.

What possible worlds might generate these two cases? In any such world, we must give up the hypothesis that the players are superhumanly rational. They must be worlds in which players sometimes make mistakes. The simplest such world arises when the mistakes are transient errors—like typos—that have no implications for mistakes that might be made in the future. In such a world, Bob still predicts that Alice will behave rationally at her second move, even though she behaved irrationally at her first move. If the counterfactuals that arise in games are always interpreted in terms of this world, then backward induction is always rational.

Perhaps the errors that a superhuman player would make are like typos, but when we apply game theory to real problems, we aren't especially interested in the errors that a superhuman player might make. We are interested in the errors that real people make when trying to cope intelligently with complex problems. Their mistakes are much more likely to be thinkos than typos. Such errors do have implications for the future. In the Chain Store Game, the fact that Alice irrationally fought at her first move may signal that she would also irrationally fight at her second move.[1] And if Bob's counterfactual is interpreted in terms of such a possible world, then the backward induction argument collapses.

To illustrate this last point, we return to the Chain Store Game of Figure 1 to see how the Nash equilibrium (*fa, oi*), though weakly dominated and not subgame perfect, might not be eliminated when we go to the limit.

Figure 2: Correlated trembles in the Chain Store Game. With probability є > 0, the Chance move of Figure 2(a) replaces Alice with a robot that always fights. Figure 2(b) shows the strategic form of the game.



(a)

	ii	io	oi	oo
aa	4-4∈, 4-4∈	4-3∈, 4+∈	10,2	10,2
af	4-4∈, 4-4∈	4-3∈, 4+∈	10,2	10,2
fa	2-2∈, 2-2∈	5,1	10,2	10,2
ff	0,0	5,1	10,2	10,2

(b)

[1] Selten repeated the game a hundred times to make this the most plausible explanation after Alice has fought many entrants in the past.

The simple trick is to expand the Chain Store Game by adding a new Chance move as in Figure 2(a). This Chance move occasionally replaces Alice with a robot player, who always fights no matter what.[2] The strategic form of Figure 2(b) shows that (*fa, oi*) is always a Nash equilibrium of the expanded game, and hence survives when we take the limit as $\epsilon \to 0$.

COMMON KNOWLEDGE OF RATIONALITY?

I think the chain-store paradox and other examples show that the claim made by Aumann (1995, 1996) and others that common knowledge of rationality implies that play will necessarily follow the backward-induction path cannot be right. The mistake is not in the proof of his theorem, but in an attempt to define rationality without saying anything about how the counterfactuals inherent in making sense of the idea are to be interpreted (Binmore 1997, 1996)). In brief, if a rational player were to make an irrational move, what would a player "with common knowledge of rationality" be entitled to deduce? If we analysts were in the habit of saying that the players in a game have a common *belief* that they are all rational—instead of common *knowledge*—we would then be forced to commit ourselves to an expanded world in which different types of players might be more or less clever when deciding how to play. In saying what was rational in the ideal world achieved by going to an appropriate limit, one would then be forced to say simultaneously what the subjective probability distribution over different possible irrational types would be in the counterfactual event that a player in our ideal world made an irrational move.

[2] For an expanded game in which all information sets are always visited with positive probability, further trembles must be added. This is why we look at (*fa, oi*) instead of (*ff, oo*), which would be eliminated if we added extra trembles in the natural way.

REFERENCES

Aumanm, R. 1995. Backward induction and common knowledge of rationality. *Games and Economic Behavior*. 8:6 -19.

Aumann, R. 1996. Reply to Binmore. *Games and Economic Behavior*. 17: 138-146.

Binmore, K. 1987. Modeling rational players I. *Economics and Philosophy*. 3: 9-55.

Binmore, K. 1996. A note on backward induction. *Games and Economic Behavior*. 17: 135-137.

Binmore, K. 1997. Rationality and backward induction. *Journal of Economic Methodology*. 4: 23-41.

Fudenberg, D. and D. Levine. 1993. Self-confirming equilibria. *Econometrica*. 61:523-546.

Kolmogrov, A. 1950. *Foundations of the Theory of Probability*. New York: Chelsea.

Lewis, David K. 1976. Counterfactuals. Oxford: Blackwell.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Selten, Reinhard and U. Leopold. 1982. Subjunctive conditionals in decision theory and game theory. In Stegmuller, Balzer, and Spohn, eds., *Studies in Economics*, Vol. 2. Berlin: Springer-Verlag.

ABOUT THE AUTHOR

Ken Binmore is Emeritus Professor of Economics at University College London. He held similar positions at the Universities of Michigan and Pennsylvania after occupying a Chair of Mathematics at the London School of Economics for many years. He is a fellow of the Econometric Society, the British Academy and the American Academy of Arts and Sciences. He is the author of some 100 published papers and 12 books, the most recent of which is *Natural Justice*, published by Oxford University Press. His research has been in game theory, bargaining, experimental economics, moral philosophy, mathematics and statistics. He has consulted widely on auction design and regulatory economics. For his part in the British $35 billion telecom auction, he was made a Commander of the British Empire.

RETURN TO SYMPOSIUM HOMEPAGE