KEN BINMORE

INTERPRETING KNOWLEDGE IN THE BACKWARD INDUCTION PROBLEM[1]

ABSTRACT

Robert Aumann argues that common knowledge of rationality implies backward induction in finite games of perfect information. I have argued that it does not. A literature now exists in which various formal arguments are offered in support of both positions. This paper argues that Aumann's claim can be justified if knowledge is suitably reinterpreted.

1. INTRODUCTION

Robert Aumann argues that common knowledge of rationality implies backward induction in finite games of perfect information (Aumann 1995, 1996a, 1996b, 1998). I have argued that it does not (Binmore 1996, 1997). A literature now exists in which various formal arguments are offered in support of both positions (for example, Bicchieri 1988; Halpern 2001; Reny 1992; Samet 1996; Stalnaker 1996), but my purpose is not to make a contribution to this literature beyond reviewing what I think the issues are. I argue instead that Aumann's claim can be justified if knowledge is suitably interpreted.

Knowledge as belief seems to be the default theoretical position both in economics and philosophy, with knowledge traditionally being defined as justified true belief. When knowledge is interpreted in this way as a special kind of belief, I continue to argue that simple counter-examples suffice to show that common knowledge of rationality does not imply backward induction. However, the hypotheses of Aumann's theorem arguably better fit an interpretation of knowledge as commitment to a model or a set of axioms.

Both varieties of knowledge seem to me already implicitly in use. For example, Socrates famously said that he knew that he knew nothing. So at least two interpretations of *knowing* must be possible if Socrates is to be absolved of

Episteme 8.3 (2011): 248–261
DOI: 10.3366/epi.2011.0020
© Edinburgh University Press
www.eupjournals.com/epi

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

contradicting himself. I think he was telling us that he was committed to the view that all his beliefs were open to doubt – including those of which he was most certain. That is to say, he had knowledge-as-commitment in the proposition that knowledge-as-belief is always open to revision. Game theorists similarly show their commitment to their knowledge of the rules of a game when they exclude illegal moves (such as moving a bishop like a knight in Chess) from a game tree, but allow for the possibility of stupid moves (like laying oneself open to Fool's Mate) even when everyone believes it certain that rational agents will not choose such moves.

Which interpretation of knowledge better fits its use in ordinary language? I think a good case can be made for knowledge-as-commitment. This paper therefore argues in favor of our always using knowledge in this sense within game theory. Knowledge-as-belief could then be referred to as certitude or belief-with-probability-one. Such a convention would concede the backward induction debate to Aumann (even in the case when rationality is interpreted in the standard Bayesian sense rather than his notion of "substantive" rationality). We could then distinguish between common knowledge of rationality and common certitude of rationality before a finite game of perfect information is played. Backward induction follows in the first case but not in the second.

2. COMMITMENT TO A MODEL

We make a commitment to a model in game theory when we specify the rules of the game we plan to study. The states of the world we plan to investigate are then normally the possible plays of the game. However the game is played, the rules of the game remain the same. Philosophers would say that the rules are the same in all possible worlds. I am suggesting that we use the same criterion for knowledge-as-commitment, even though we then end up with a notion in which Bob may know something that is neither true nor justified.

Wittgenstein (1969) argues that something close to what I am calling knowledge-as-commitment is unavoidable, but I do not follow him much further. My more mundane approach goes back to the basics of orthodox decision theory, in which preferences and beliefs are deduced from consistent choice behavior. I argue that the same should be true of our conception of knowledge. My suggested definition of knowledge-as-commitment therefore says that Bob knows something if he acts as though it were true in all possible worlds that might arise within whatever model he is using.

With this definition, nothing can happen to shake Bob's commitment to what he knows. This accords with the convention in real life according to which we say that Bob was mistaken in saying that he knew something if he later changes his mind about it. An abuse of the same convention leads people to insist that they know things in order to advertise their commitment to whatever they may be claiming. For example, the philosopher Moore knew that what we commonly call a tree is

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

indeed a tree. Kant knew that space is Euclidean. Ronald Reagan knew in his heart that man is good. Keats knew that truth is beauty. And as an example of religious faith, Handel knew that his redeemer liveth.

However, the usage I am talking about is better illustrated in a scientific context. When physicists are asked to explain their ontology, they do not usually imagine that they are being asked a metaphysical question about the fundamental nature of reality. They simply reply by stating the propositions that are taken to be axiomatic in their models. Their faith in these axioms may only be provisional, but commitment to a model does not require the kind of faith that martyrs die for.

I suspect that it is an unconscious acceptance of such an interpretation of knowledge that leads game theorists to accept the same axioms for their knowledge operator $\mathcal{K}$ as philosophers commonly demand for their necessity operator $\Box$, which is used to prefix a proposition that holds in all possible worlds (section 5).

Contradicting knowledge? What if a possible world occurs that embodies an in-your-face contradiction of something that Bob knows? Knowledge-as-commitment requires ignoring the contradiction and continuing to uphold what was previously held to be known. An analogy is to be found when mathematicians are led to a contradiction after adjoining a new proposition to an axiom system. One could deal with the problem by throwing away one or more axioms, but it is actually the new proposition that is said to be refuted.

It seems to me that we often similarly dismiss data that does not match our preconceptions. An extreme example is provided by the colleague of Galileo who supposedly refused to look through Galileo's telescope lest he see the moons of Jupiter. Such behavior is much more common than we usually like to admit. Rather than confess to having been mistaken, people invent an extraneous reason to explain away inconvenient data. Perhaps it was garbled in transmission. Perhaps the experiment from which it was derived was bungled. Perhaps a hand trembled when it reached for a strategy button, and pressed the wrong one by mistake. Kuhn (1962) includes making up such excuses as part of the paradigm of "normal science".

Such behavior need not be intellectually dishonest, because few models (if any) succeed in taking account of everything that might matter. Off-model explanations of apparent contradictions are therefore sometimes a practical necessity. However, it is important that such off-model explanations should not have any implications for the future. Otherwise, the knowledge that is being insulated from a current challenge would be laid open to future challenges and hence not properly counted as knowledge. For this reason, the error distributions in refinements of Nash equilibrium like Selten's (1975) trembling-hand equilibrium are taken to be independent of past error distributions. Philosophers similarly urge interpreting counterfactuals so as to minimize the extent to which the basic model needs to be modified to accommodate the counterfactual phenomena (Lewis 1973).

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

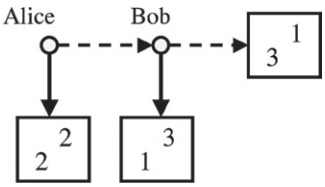


Figure 1. A simple example. Alice's utilities are written in the southwest of each payoff box and Bob's in the northeast. The firmly drawn arrows indicate the players' backward induction choices. The reasoning is that Bob would play *down* if his decision node were reached because he would then gain a payoff of 3 rather than 1. Alice predicts Bob's choice and so plays *down* herself to gain a payoff of 2 rather than 1.

3. BACKWARD INDUCTION

What I have to say on this subject is that some of the assumptions that Aumann (1995) builds into his definition of substantive rationality (or Sobel (1994, ch. 16) into his definition of resilient rationality) are better subsumed within a conception of knowledge as commitment to a model.

Aumann (1995, sect. 4b) emphasizes that, in his model, "the rationality of a player at a vertex v of a game is defined in terms of what happens at vertices *after* v." A footnote warns that this requirement is open to misunderstanding, but fails to dispel the widespread belief that Aumann is talking about common *certitude* of *Bayesian* rationality before the game is played.[2]

The game of figure 1 illustrates why this cannot be the case. The use of backward induction in this game will result in Alice playing down at the first move. For this move to be rational, Alice must believe that Bob would play down (with sufficiently high probability) should he get to move. A more telling example will be given in section 4, but even in this very simple game, common certitude of Bayesian rationality before the game is played does not logically imply backward induction. If the claim were correct and Bob were to get to play, Alice would have acted irrationally and hence contradicted the hypothesis that everyone is rational. Since anything follows from a contradiction, we are then free to attribute any behavior to Bob. In particular, Aumann is free to hypothesize that Alice will believe that if she were to play across, then Bob would behave rationally by playing down. Aumann seeks to do so by expanding the definition of rationality into a criterion that applies independently at *every* vertex v, including those vertices that are not reached when backward induction is used. To quote Aumann: "If we demand rationality only at vertices that are actually reached, the theorem fails already in the simplest cases" (1995, 12).

In brief, Aumann's definition of (substantive) rationality is constructed so that his underlying hypothesis would be maintained no matter how often it might be contradicted by the evidence that becomes available as the game is played. I think

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

that such assumptions about counterfactuals, whether made implicitly or explicitly, do not belong in a definition of rationality.[3] Clarity demands that assumptions about what the players would believe if things that are not going to happen were to happen need to be made somewhere else.

The suggestion made in this paper is that we avoid writing Aumann's (implicit) assumptions on counterfactuals into a definition that expands what rationality had hitherto been taken to mean in game theory. They should be written instead into a new definition of knowledge that makes explicit the common-sense idea that you do not really know something today if it is possible that you might change your mind about it tomorrow. Such knowledge cannot be updated. Updating is something one does with beliefs within the context of whatever model one is committed to.

4. BELIEFS WITHIN A MODEL

In the terminology I advocate, one *knows* things about the model or world that underlies an analysis. One has *beliefs* about the various states of the world that may arise within the model. If the model is sufficiently small for Savage's (1954) axioms to apply,[4] then Bob's beliefs within the model can be described using subjective probability measures. The evolution of his beliefs over time can be tracked using Bayes' rule, provided that it is never necessary to update on an event previously assigned probability zero.

4.1. Chain-Store Paradox

Reinhard Selten's (1978) Chain-Store paradox is the classic example of a game that shows that backward induction cannot always be right because it ignores the necessity of interpreting the counterfactuals that keep players on the equilibrium path. In the story that goes with the game, Alice's chain of stores operates in 100 towns. Bob contemplates setting up a rival stores successively in each town. If Bob sets up store in a town, Alice can acquiesce in his entry by sharing the market, or fight a price war in that town.

Figure 2(a) shows what the game would be if there were only one town. Figure 2(b) is a simplified version of what the game would be if there were two towns (Binmore 2007, 409). The simplification is that if Bob chooses to stay out of the first town, it is assumed that he necessarily stays out of the second town. Similarly, if Alice concedes in the first town after Bob enters, then it is assumed that Bob necessarily enters the second town, and Alice again concedes. This simplified version will be used to study the viability of backward induction in the Chain-Store paradox.[5]

The difficulty with backward induction arises in its most acute form when Bob plans what choice to make at his second move in the version of the Chain-Store paradox of figure 2(b). If backward induction is correct, then Alice will

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

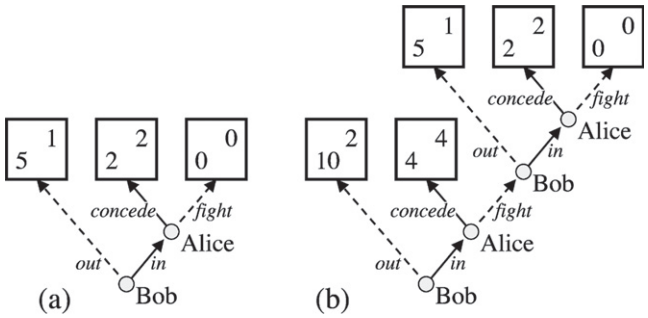


Figure 2. Chain-Store Paradox. The game on the left would be played if there were only one town in the Chain-Store paradox. The game on the right is a simplified version of the game with two towns in which it is assumed that payoffs received at different times can be added without any discounting. The simplification is that if Alice concedes in the first town then she must concede in the second town. Similarly, if Bob does not contest in the first town then he cannot contest in the second town. The firmly drawn arrows indicate the backward induction choices.

concede at her first move. Since Bob's second move cannot be reached unless Alice fights at her first move, Bob therefore has a counterfactual to evaluate. Whatever assumptions led him to the conclusion that backward induction is correct have just been refuted by Alice's last move. He therefore cannot rely on these assumptions to predict Alice's next move. So how should he evaluate the probability that Alice will fight at her second move given the supposedly counterfactual event that she fought at her first move?

Much has been written on the interpretation of counterfactuals but the formalisms that have been proposed seldom seem helpful in the kind of problems that arise almost immediately in game theory. In particular, Lewis's *Counterfactuals* (1973) tells us to interpret a counterfactual event as though it had occurred in the nearest "possible world" in which its occurrence would not be counterfactual. Kolmogorov (1950) says something similar about conditioning on zero probability events, but gives examples that illustrate the ambiguity implicit in such statements. The problem is that nothing says which possible world is "nearest". In game theory, there are always multiple candidates.

Selten and Leopold (1982) argue that the way to deal with counterfactuals in game theory is to elaborate the game so that a counterfactual in the original game ceases to be counterfactual in the elaborated game. That is to say, they argue in favor of formalizing the kind of off-model stories used to "explain away" events that contradict a treasured model. They then follow Kolmogorov in suggesting that the newly invented parameters of the elaborated game be allowed to converge to zero, so that the elaborated game converges on the original game. One then interprets a counterfactual in terms of the limiting form of whatever factual statement in the elaborated game corresponds to the counterfactual in the original game. This approach is standard within game theory but cannot easily be

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

accommodated within any formalism intended to be generally applicable because it relies on modeling the *context* in which a particular game has arisen.

One way of elaborating a game is to follow all moves with independent chance moves (or trembling hands) that assign small probabilities to each possible alternative to the choice a player actually made at that move. Counterfactual sequences of choices cannot arise in such an elaborated game, which may therefore serve as a means of explaining counterfactuals in the original game. Selten (1975) has worked out the consequences when trembling hands always assign each alternative the same small probability. When the probabilities of a tremble are allowed to become vanishingly small, the resulting "trembling-hand equilibria" are (almost) the same as the equilibria obtained through backward induction in finite games of perfect information. Of course, one need not assume that the probabilities of trembles are all equal. Myerson (1991) is thereby led to an equilibrium refinement called "proper equilibrium". Other assumptions will lead to other refinements. Nor need one assume that trembles are independent. If one is sufficiently generous in what can be counted as a tremble, Fudenberg, Kreps, and Levine (1988) have shown that there is always some way to interpret counterfactuals so as to justify any Nash equilibrium at all.

It is empty to ask what are the "correct" trembles to use in interpreting the counterfactuals that might arise in a game. As observed earlier, I disagree with those philosophers who believe that a useful answer can be found by applying some general algorithm to the formal expression of a game. I share the view of those philosophers who argue that the place to look for useful answers is in the context in which a counterfactual arises. When contemplating Aumann's (1995) claim that common knowledge of rationality implies backward-induction play in finite games of perfect information, one therefore needs to adopt some view about the context under consideration. With knowledge-as-commitment, the context is laid on the table.[6] But with knowledge-as-belief (as rationality is usually interpreted in game theory), the context is open for debate.

However, interpreting knowledge about the rationality of the players as a kind of belief tells us something in itself about the appropriate context. Suppose that all the players believe that they are all Bayesian rational at the outset of the game. According to orthodox Bayesian theory, the implication is that there is a probability space S whose states specify (among other things) whether or not the players are rational. Some states must therefore describe situations in which one or more of the players are irrational. But there are many possible ways of being irrational. To specify the state space fully, we would therefore need to itemize all the different ways in which a player might be irrational. To model the players' beliefs about the rationality of their opponents, we then need to elaborate the game by imagining an opening chance move that selects a state from S. With knowledge-as-belief, the players then believe with probability one that the selected state will have the property that all the players are rational.

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

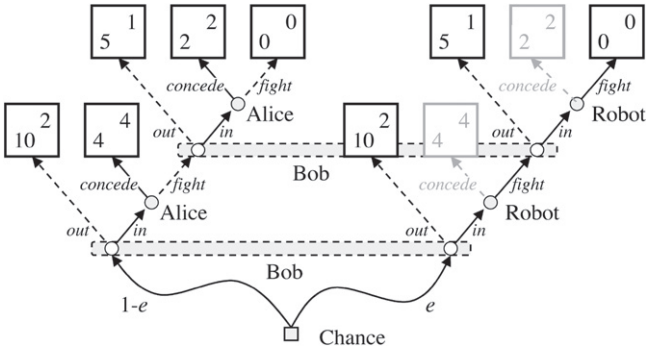


Figure 3. Introducing the possibility of some irrationality. The game of figure 2(b) has been perturbed by introducing a chance move that may replace Alice by a robot that always fights no matter what. This event occurs with some small probability $e > 0$. (Since the robot is not able to concede, the results of a concession by the robot have been de-emphasized to indicate that this choice is not part of the perturbed game.) Information sets have been introduced into the perturbed game to indicate that Bob is unaware of the outcome of the chance move. The firmly drawn arrows show the choices that correspond to applying backward induction to the unperturbed game of figure 2(a).

Given this set-up, how are counterfactuals to be interpreted? An obvious approach touted in (Binmore 1987) is to regard the players' common certitude that they are all rational as a limiting case of the situation in which the chance move that selects a state from S does not select irrational types only with zero probability.[7] With some small positive probability, a state will then be selected from S in which one or more players are irrational in one or more of the many ways in which it is possible to be irrational.[8] If S is sufficiently rich, all counterfactual events can then be explained away by postulating sufficiently exotic forms of irrationality.

The Chain-Store paradox was introduced at the beginning of this section in order to illustrate this point. The original game that needs to be elaborated is given in figure 2(b). The elaboration shown in figure 3 consists of a perturbation that selects an irrational player (or robot) with some small probability $e > 0$ to occupy the role of Alice.[9] In this example, the only kind of irrationality considered has the robot always choosing to fight no matter what. Introducing other kinds of irrationality would sometimes lead to different conclusions, but we need only one counter-example to cast doubt on how Aumann's claim has been interpreted by people less sophisticated than Aumann himself.

The firmly drawn arrows in figure 3 show the choices that would be made by Alice and Bob in the elaborated game if they played as backward induction requires in the original game of figure 2(b). The choice attributed to Bob at his second information set is then not optimal. The reason is that Bob's arrival at his second information set would trigger a reappraisal of the probability that his opponent is Alice. His prior probability e that his opponent is the robot would need to be

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

updated to a posterior probability of 1 given that Bob is certain that Alice will concede at her first move (because such a concession by his opponent would make it impossible for Bob's second information set to be reached).

In fact, the only Nash equilibria in the elaborated game require Alice to plan to fight in the first town and for Bob to stay out of the market in the first town. Since this is true for all $e > 0$, it remains true in the limiting case as $e \to 0$. With this interpretation of the counterfactuals that arise in the original game of figure 2(b), it follows that backward induction in the original game is not rational (in the ordinary Bayesian sense of rational, as opposed to Aumann's redefinition of the term).[10] Of course, other interpretations of the counterfactuals may lead to other conclusions. In particular, Aumann's claim about common knowledge of rationality implying backward induction is indeed true with his implicit interpretation of how counterfactuals should be interpreted. My complaint is only that it needs to be recognized that Aumann is indeed making assumptions about the interpretation of counterfactuals (Halpern 2001).

5. PROPERTIES OF KNOWLEDGE AS COMMITMENT

Although game theorists are largely unaware of the fact, the standard assumptions about knowledge necessary to justify the way they model information sets in games are captured by the axioms listed in figure 4 (Binmore 2007, 356). The axioms (K0)–(K4) that govern the use of the knowledge operator K are those of the modal logic S5. However, in accordance with the traditions of economic decision theory, they are expressed in set-theoretic terms. Thus Ω represents the underlying state space. An event E is a subset of Ω. The set KE is the event that Bob knows that E has occurred. The event PE is the event that Bob regards it possible that E has occurred.

(K0)	$\mathcal{K}\Omega = \Omega$	(P0)	$\mathcal{P}\emptyset = \emptyset$
(K1)	$\mathcal{K}(E \cap F) = \mathcal{K}E \cap \mathcal{K}F$	(P1)	$\mathcal{P}(E \cup F) = \mathcal{P}E \cup \mathcal{P}F$
(K2)	$\mathcal{K}E \subseteq E$	(P2)	$\mathcal{P}E \supseteq E$
(K3)	$\mathcal{K}E \subseteq \mathcal{K}^2E$	(P3)	$\mathcal{P}E \supseteq \mathcal{P}^2E$
(K4)	$\mathcal{P}E \subseteq \mathcal{K}\mathcal{P}E$	(P4)	$\mathcal{K}E \supseteq \mathcal{P}\mathcal{K}E$

Figure 4. Knowledge and possibility. The most controversial of these axioms is (K4). It can be rewritten as $(\sim\mathcal{K})^2E \subseteq \mathcal{K}E$, which is interpreted as meaning that if Bob doesn't know that he doesn't know E, then he knows E.

The possibility operator $\mathcal{P}$ may be defined by $\mathcal{P} = \sim\mathcal{K}\sim E$, in which case the axioms (P0)–(P4) are equivalent to (K0)–(K4). A possibility set $P(\omega)$ is the set of all states that Bob would regard as possible if the true state were ω. Such a possibility set is related to the possibility operator $\mathcal{P}$ by the equivalence

$$x \in P(y) \Leftrightarrow y \in \mathcal{P}\{x\}.$$

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

In game theory, it is taken for granted that such possibility sets partition Ω, since information sets (such as those that appear in figure 3) are simply elaborations of possibility sets that take account of the time at which information is received. This partitioning property is equivalent to the axioms (K0)–(K4) with (K2) excluded.

How reasonable the axioms seem depends on how the knowledge operator is interpreted. For example, it appears to be generally accepted that the very strong axiom (K4) is unreasonable if knowledge is interpreted as awareness. On the other hand, (K4) would seem to make sense in the small world shared by the players of a game if they are assumed to have considered all possible implications of what can happen if they take the rules of their game to be axiomatic. But one then runs into trouble with (K2) because although Alice and Bob may take it to be axiomatic that the rules of their game will not be broken, it is not in fact true that the rules of games like Chess are unbreakable.

My proposal is that we escape such difficulties by adopting the same attitude that modern mathematicians take to axiom systems. Insofar as they "believe" in their axioms, their belief cannot be said to be justified true belief. The axioms of Euclidean geometry, for example, are now so far from being regarded as self-evident that they are thought to be false if interpreted as describing the properties of the world in which we live. As for justification, one would not speak of axioms at all if any kind of *formal* justification were going to be offered. It is for this reason that Bertrand Russell (1917/2007) said that mathematics could be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true. Mathematicians say that a statement is an axiom to indicate that they are committed to behaving as if it were true in the context under study. When arguing by contradiction, they even commit themselves temporarily in this way to treating statements as true that they plan to refute.

In the wider scientific world, one commonly speaks of a model rather than an axiom system, but the same considerations apply. For example, I can recall being taught how fluid dynamics would work in a two-dimensional world, and what the orbits of the planets would be if gravity obeyed an inverse cube law. In game theory, one often similarly commits oneself to assumptions that fail all the tests for justified true belief. A case in point is the standard assumption that there is common knowledge of rationality. As I have argued elsewhere, it is very hard for anything whatever to become common knowledge in the real world, let alone such an elusive concept as the rationality of a player (Binmore 2008). However, I agree with Aumann that it is worthwhile to study the implications of making it axiomatic that there is common knowledge of rationality for the same reasons that physicists think it worthwhile to study two-dimensional flows. Where we disagree is on how rationality and knowledge can most usefully be defined.

In the small world of a game, the knowledge axioms of figure 4 other than (K2) can be defended in the case of knowledge-as-commitment with the same informal arguments as those offered for knowledge-as-belief. Of course, one steps outside

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

the small world of a game when one starts modeling the players themselves, as is necessary when assuming that there is common knowledge of rationality among the players. However, Bayesians sometimes argue that one can reduce this larger world to a small world (in Savage's sense) by postulating a chance move that selects the players' levels of rationality before the game is played (as in the discussion of the Chain-Store paradox of section 4.1).

The axiom (K2) is more difficult, as we can no longer interpret it as meaning that knowing E implies that E is true in some absolute or metaphysical sense. We need to substitute the requirement that the player in question will act *as though* E were true.

A statement that is true in an absolute sense is presumably invulnerable to contradiction, but what of a statement that is only held to be true by some fallible human being? A very determined Bayesian might express this problem by writing

$$\text{(1)} \qquad \text{prob}(E|\sim E) = \frac{\text{prob}(E \cap \sim E)}{\text{prob}(\sim E)} = \frac{\text{prob}(\emptyset)}{\text{prob}(\sim E)},$$

from which it follows that either $\text{prob}(E|\sim E) = 0$ or $\text{prob}(\sim E) = 0$. As explained in section 2, a person who is treating E as axiomatic will opt for the second of these alternatives. Observing the zero-probability event $\sim E$ will then not be treated as a contradiction of E that refutes Bob's underlying model, but as a counterfactual event that needs to be interpreted like other counterfactual events in a manner dependent on the context.

Updating parameters using Bayes' Rule. Replacing knowledge-as-belief by knowledge-as-commitment allows the problem of scientific induction to be freed from the (naive) claim that it can be reduced to the kind of Bayesian updating that follows from honoring Savage's consistency axioms (Savage 1954; Binmore 2009). Scientific revolutions arise when scientists find that the tensions involved in explaining away anomalies in the data with off-model stories become too great to allow them to maintain their commitment to a particular model. They then throw away their old model and adopt a new model – freely admitting as they do so that they are being inconsistent. Bayesian updating has no place in such a revolution. It needs to be seen instead merely as a humdrum updating of Bob's beliefs about the parameters of a model to whose structure he is currently committed.

How does knowledge-as-belief or belief-with-probability-one fit within such a scheme? On this subject, I share the view of those Bayesians who insist that one should never allow subjective probabilities to go all the way to zero or one. Otherwise any attempt by Bob to adjust his system of beliefs into consistency with new data would fail when, as in equation (1), he found himself trying to condition on a zero probability event. This is not to say that models in which

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

non-trivial events are accorded a subjective probability of one may not often be useful; only that they should be regarded as limiting cases of models in which extreme probabilities do not appear.

6. CONCLUSION

Knowledge as commitment to a model is impervious to revision within the model in question. It then becomes trivially true that common knowledge of rationality implies backward induction. No history of play can disturb the initial assumption that there is common knowledge of rationality, because knowledge is not subject to revision. In predicting the behavior of the players in any subgame, we can therefore count on this hypothesis being available. It may be that some of these subgames could not be reached without Alice repeatedly choosing actions that game theorists would classify as irrational, but if Bob *knows* that Alice is rational, her history of irrational behavior is irrelevant. He will continue to act as though it were true that she is rational in all possible worlds that may arise.

Accepting this proposed resolution of the backward induction problem has implications for the value of the various formalisms that have been brought to bear on the problem in the past. It seems to me that they have served only to kick up dust that has obscured what is essentially a trivial issue. A lesson for which there will be few takers is that mathematics is often best left on the shelf until the philosophy of a problem has been properly sorted out.

REFERENCES

Aumann, R. 1995. "Backward Induction and Common Knowledge of Rationality." *Games and Economic Behavior* 8: 6–19.

Aumann, R. 1996a. "Reply to Binmore." *Games and Economic Behavior* 17: 138–46.

Aumann, R. 1996b. "Reply to Binmore and Samuelson." In K. J. Arrow et al. (eds.), *The Rational Foundations of Economic Behavior*, pp. 130–1. London and Basingstoke: Macmillan.

Aumann, R. 1998. "On the Centipede Game." *Games and Economic Behavior* 23: 97–105.

Bicchieri, C. 1988. "Common Knowledge and Backwards Induction: A Solution to the Paradox." In M. Vardi (ed.), *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 381–93. Los Altos, CA: Morgan Kaufmann.

Binmore, K. 1987. "Modeling Rational Players II." *Economics and Philosophy* 4: 179–214.

Binmore, K. 1996. "A Note on Backward Induction." *Games and Economic Behavior* 17: 135–7.

Binmore, K. 1997. "Rationality and Backward Induction." *Journal of Economic Methodology* 4: 23–41.

Binmore, K. 2007. *Playing for Real*. New York: Oxford University Press.

Binmore, K. 2008. "Do Conventions Need to Be Common Knowledge?" *Topoi* 27: 17–27.

Binmore, K. 2009. *Rational Decisions*. Princeton: Princeton University Press.

Fudenberg, D, D. Kreps, and D. Levine. 1988. "On the Robustness of Equilibrium Refinements." *Journal of Economic Theory* 44: 354–80.

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

Halpern, J. 2001. "Substantive Reasoning and Backward Induction." *Games and Economic Behavior* 37: 425–35.
Kolmogorov, A. 1950. *Foundations of the Theory of Probability*. New York: Chelsea.
Kreps, D, P. Milgrom, J. Roberts, and R. Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* 27: 245–52.
Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
Lewis, D. 1973. *Counterfactuals*. Oxford: Blackwell.
Myerson, R. 1991. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.
Reny, P. 1992. "Rationality in Extensive Form Games." *Journal of Economic Perspectives* 6: 103–18.
Russell, B. 1917/2007. *Mysticism and Logic*. Nottingham: Spokesman.
Samet, D. 1996. "Hypothetical Knowledge and Games with Perfect Information." *Games and Economic Behavior* 17: 230–51.
Savage, L. 1954. *The Foundations of Statistics*. New York: Wiley.
Selten, R. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive-Games." *International Journal of Game Theory* 4: 25–55.
Selten, R. 1978. "The Chain-Store Paradox." *Theory and Decision* 9: 127–59.
Selten, R and U. Leopold. 1982. "Subjunctive Conditionals in Decision Theory and Game Theory." In W. Stegmüller, W. Balzer, and W. Spohn (eds.), *Philosophy of Economics: Proceedings, Munich, July 1981*. Berlin: Springer-Verlag.
Sobel, J. 1994. *Taking Chances*. Cambridge: Cambridge University Press.
Stalnaker, R. 1996. "Knowledge, Belief and Counterfactual Reasoning in Games." *Economics and Philosophy* 12: 133–63.
Wittgenstein, L. 1969. *On Certainty*. G. Anscombe and G. von Wright (eds.). D. Paul and G. Anscombe (trans.). Oxford: Basil Blackwell.

NOTES

1 I am grateful to the British Arts and Humanities Council for funding this work through grant AH/F017502/1 and to Robert Aumann and Brian Skyrms for their helpful advice.
2 If Bob is Bayesian rational, he maximizes his expected payoff relative to a subjective probability distribution that summarizes his beliefs. He updates his beliefs when new information is received using Bayes' rule.
3 Actually, even with Aumann's definition, the problem of interpreting counterfactuals does not disappear. Indeed, Stalnaker (1996) interprets the relevant counterfactuals in a manner that renders Aumann's conclusion false. Halpern (2001) clarifies the issues very successfully, but the only essential point here is that Aumann does make implicit assumptions about counterfactuals (although these assumptions seem entirely reasonable to me within the context in which the backward induction problem arises and Aumann's definition of substantive rationality).
4 The important fact that Savage restricted his theory of decision under uncertainty to what he called small worlds is largely overlooked (Binmore 2009).

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press

5 In using a version of the story with only two towns, one loses some of the intuitive force of the original story in which one is asked how likely it would seem to Bob that Alice would concede in the 100th town after having fought in 99 preceding towns.

6 One might think that the assumption of common knowledge would then be too strong to be very useful in practice, but this is irrelevant to Aumann's claim.

7 More generally, I think that one should never discuss belief with probability one except in the context of an idealization obtained by allowing non-extremal probabilities to become extremal.

8 If common knowledge of the rules of the game is being understood in the commitment sense rather than the belief sense, one would not then wish to follow Selten by inserting trembles into the *rules* of an elaborated game. Trembling tendencies or failures of memory and the like would be better included as being part of the description of possible types of irrational players.

9 There is nothing new in such a modeling approach. The best known example is the famous Gang of Four paper in which a small probability that the opponent might (irrationally) be committed to playing tit-for-tat in the finitely repeated Prisoners' Dilemma results in cooperation among the rational players until near the end of the game (Kreps et al. 1982).

10 Note that if the story that leads to our elaboration of the original game is retold in terms of Selten's trembling hands, then the trembles need to be *correlated* rather than independent in the new story. Otherwise Bob would have no grounds for deducing from the fact that his opponent fought in the past that she is likely to fight in the future.

Ken Binmore is a mathematician turned economist and philosopher who is currently a Visiting Professor of Economics at the University of Bristol and a Visiting Professor of Philosophy at the London School of Economics. He currently works mostly in experimental economics, evolutionary biology, and moral philosophy. Recent books include *Natural Justice* (OUP), *Does Game Theory Work?* (MIT Press), *A Very Short Introduction to Game Theory* (OUP), and *Rational Decisions* (Princeton).

https://doi.org/10.3366/epi.2011.0020 Published online by Cambridge University Press