Can Knowledge Be Justified True Belief?

Ken Binmore
Economics Department
University College London
Gower Street
London WC1E 6BT, UK

Abstract: Knowledge was traditionally held to be justified true belief. This paper examines the implications of maintaining this view if justication is interpreted algorithmically. It is argued that if we move sufficiently far from the small worlds to which Bayesian decision theory properly applies, we can steer between the rock of fallibilism and the whirlpool of skepticism only by explicitly building into our framing of the underlying decision problem the possibility that its attempt to describe the world is inadequate.

Can Knowledge Be Justified True Belief?[1]

by Ken Binmore

1 Fallibilism or Skepticism?

The view that knowledge can usefully be interpreted as justified true belief has fallen into disfavor in recent times. David Lewis [8] observes that the use of such a definition seems to require an apparently impossible choice between the rock of fallibilism and the whirlpool of skepticism, but that we can—just barely—escape both perils by steering with care. This paper offers a more radical defense of the same conclusion.

A standard objection to the traditional definition has been voiced by Gettier [7]. In a bowdlerized version of his story, Boris and Vladimir have both proposed marriage to the beautiful Olga. She blushes with pleasure when Vladimir pays her compliments, but seems not to remember Boris at all. Boris is not surprised, because he is poor and Vladimir is rich. He thinks his knowledge of the world justifies his believing that Olga will marry money. When Olga surprises Boris by accepting his proposal, his belief turns out to be true because, unknown to anyone, a long-lost uncle has died and left Boris a fortune. But do we want to say that Boris therefore *knew* that Olga would marry a rich man?

The traditional definition can be defended against such attacks by challenging the standard of justification that is employed. In our Russian story, Boris should perhaps have been more realistic about his own inexperience in matters of the heart, and sought advice from an agony aunt. He would then have learned that beautiful maidens sometimes pretend indifference with a view to fanning the flames of a favored suitor's ardor.

This paper avoids such disputes over what counts as adequate justification by treating the justification process as algorithmic. The general problems that arise in treating knowledge algorithmically are surveyed in Binmore and Shin [6]. This paper confines its attention to investigating the formal implications of maintaining that knowledge must simultaneously be justified and true. The findings depend on the manner in which the underlying decision problem is framed.

In formulating what is nowadays called Bayesian decision theory, Leonard Savage [9] distinguished between large and small worlds. His *Foundations of Statistics*

[1]I am grateful to Alan Hajek for his valuable comments.

says at one point that it would be "ridiculous" and at another that it would be "preposterous" to apply his theory in a large world, but no formal criteria are offered to distinguish between small-world decision problems and large-world decision problems (Binmore [3]).

This dimly recognized distinction between large and small worlds in decision theory echoes a distinction in proof theory made precise by Gödel. A mathematical system large (or complex) enough to include arithmetic cannot be simultaneously consistent and complete. This paper adapts the halting argument for Turing machines in defending the claim that decision theory needs to recognize a similar distinction.

In a context sufficiently far removed from the small worlds to which Bayesian decision theory properly applies, the knowledge assumptions that the theory implicitly takes for granted can no longer be sustained. But we can still steer between the rock of fallibilism and the whirlpool of skepticism by explicitly building into our framing of the underlying decision problem the possibility that any such framing may fail to capture something significant.

2 Small World Assumptions

Bayesian decision theory takes for granted that a decision-maker's knowledge at any time partitions the universe of discourse into a collection of disjoint possibility sets. The partitioning property of these possibility sets is then inherited by the information sets that Von Neumann introduced into game theory (Binmore [5, p.454]). An assumption of 'perfect recall' is then usually made to ensure that a player's knowledge partition is always a refinement of his previous partitions. This section reviews the linkage between such knowledge partitions and the idea of a knowledge operator.

We identify an event E with a subset of a given universe of discourse denoted by Ω. The event in which Boris knows that E has occurred is denoted by $\mathcal{K}E$, where $\mathcal{K}$ is his knowledge operator. The event in which Boris thinks it possible that E has occurred is denoted by $\mathcal{P}E$, where $\mathcal{P}$ is his possibility operator.

If we make the identification $\mathcal{P} = {\sim}\mathcal{K}{\sim}$, then we establish a duality between $\mathcal{K}$ and $\mathcal{P}$. Either of the following lists will then serve as a rendering of the requirements of the modal logic S–5:

(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)	$\mathcal{K}E \supseteq \mathcal{P}\mathcal{K}E$

The "infallibility" axiom can be taken to be either (K2) or (P2). The seemingly innocent (K0) and (P0) will be called "completeness" axioms.

These ideas are linked with knowledge partitions by defining the possibility set $P(\omega)$ to be the set of states that Boris thinks is possible when the true state of the world is ω. In Bayesian decision theory, the minimal requirements for such possibility sets are usually taken to be:

(Q0) The collection $\{P(\omega) : \omega \in \Omega\}$ partitions Ω
(Q1) $\omega \in P(w)$

The second of these assumptions is the "infallibility" requirement.

To establish an equivalence between the two approaches, it is only necessary to define $\mathcal{P}$ and P in terms of each other using the formula:

$$\omega \in \mathcal{P}E \;\Leftrightarrow\; P(\omega) \cap E \neq \emptyset\,. \tag{1}$$

With this definition, (P0)–(P3) can be deduced from (Q1)–(Q2) and vice-versa. However, the role of the completeness axiom (P0) is peripheral. If we dispense with (P0) and redefine both (P1)–(P3) and (1) so that they apply only to non-empty events, then (new P1)–(new P3) are equivalent to (Q1)–(Q2).

It is significant that (P0) can be eliminated, because the result of the next section can be regarded as saying that (P0) and (P2) cannot both hold in a large enough world when the possibility operator is algorithmic.

3 Justification

The process of justification is abstracted away in the previous section. It is understood to be somehow built into the knowledge or possibility operators. We now unpack this black box by postulating that justification is actually carried out by a 'Leibniz engine' J that makes judgements on what events are possible.

The assertion that justification is algorithmic is interpreted to mean that J is a Turing machine that sometimes answers NO when asked questions that begin:

Is it possible that ...?

Issues of timing are obviously relevant here. How long does one wait for an answer before acting? Such timing problems are idealized away by assuming that Boris is able to wait any finite number of periods for an answer.

As in the Turing halting problem, we suppose that $[\,N\,]$ is some question about the Turing machine N. We then take $\{M\}$ to be the question:

Is it possible that M answers NO to $[\,M\,]$?

Let T be the Turing machine that outputs $[\,x\,]$ on receiving the input $\{x\}$, and employ the Turing machine $I = JT$ that first runs an input through T, and then runs the output of T through the justification machine J. Then the Turing machine I responds to $[\,M\,]$ as J responds to $\{M\}$.

An event E is now defined to be the set of states in which I responds to $[\,I\,]$ with NO. We then have the following equivalences:

$$\begin{aligned}\omega \in E \quad &\Leftrightarrow \quad I \text{ responds to } [\,I\,] \text{ with NO} \\ &\Leftrightarrow \quad J \text{ reports it to be impossible that } I \text{ responds to } [\,I\,] \text{ with NO} \\ &\Leftrightarrow \quad \omega \notin {\sim}\mathcal{P}E\end{aligned}$$

It follows from (P2) that

$$\sim\mathcal{P}E \;=\; E \;\subseteq\; \mathcal{P}E\,.$$

This identity only holds when $\mathcal{P}E = \Omega$. Since $E = \sim\mathcal{P}E$, it follows that $E = \emptyset$, and so $\mathcal{P}\,\emptyset = \Omega$. That is to say, we are led to the following apparent contradiction:

> **Proposition.** If the states in Ω are sufficiently widely construed and knowledge is algorithmic, then infallibility implies that the decision-maker always thinks it possible that nothing will happen.

If one seeks to maintain (P0) or (K0) in a world large enough for our use of the Turing argument to make sense, this proposition puts paid to Lewis's attempt to steer between the rock of fallibilism and the whirlpool of skepticism . But why should we hang on to these hard-to-interpret completeness axioms in a large world?

4 What is an event?

I think the apparent contradiction built into the preceding proposition signals a failure of the model to capture the extent to which familiar assumptions from small-world decision theories need to be modified when moving to a large world.

For example, we think of an event E as having occurred if the true state ω of the world has whatever property defines E. But how do we determine whether ω has this property?

If we are committed to an algorithmic approach, we need an algorithmic procedure for the defining property P of each event E. This procedure can then be used to interrogate ω with a view to getting a YES or NO answer to the question:

Does ω have property P?

We can then say that E has occurred when we get the answer YES, and that $\sim E$ has occurred when we get the answer NO.

But in a sufficiently large world, there will necessarily be properties for which the relevant algorithm sometimes will not halt. Our methodology will then classify ω as belonging neither to E nor to $\sim E$. Our inadequate formalism then forces us to place ω in $\emptyset$—although we can no longer interpret this as the set with no elements.

A more satisfactory analysis would perhaps appeal to some appropriate version of constructivist or intuitionistic logic,[2] but I hope the preceding remarks will at least make it plausible that we can sustain the conclusion of the proposition of the previous section in a large world. To say that $\mathcal{P}\emptyset = \Omega$ can be interpreted to mean that Boris necessarily thinks it possible that the true state of the world will remain unclassified according to any of the properties recognized by his algorithmic classification system.

5 All-encompassing worlds

Robert Aumann [1] has pioneered an approach to the foundations of game theory in which the states of the world in the model are to be thought of as encompassing absolutely everything that could conceivably happen—including Boris's states of mind and behavior. I have used Gödelian arguments elsewhere to criticize Aumann's use of Bayesian decision theory in such a large-world context (Binmore [2, 4]). But what if Aumann's visionary approach is employed, using a decision theory that is suited to large-world applications?

On this subject, I shall only point out that the argument of Section 3 survives allowing the justification machine J to depend on the true state of the world. Boris's justification algorithm is then one of the many properties of whatever the true state ω turns out to be. When Boris interrogates ω, he will then sometimes be asking a question about the workings of his own cognitive processes.

The non-halting argument of Section 3 is based on precisely this self-referential possibility. The argument can therefore be seen as another telling of the tale that Boris cannot operate an algorithmic model that is always successful in predicting the workings of his own mind. Still less can Boris operate an algorithmic model that is always successful in predicting what the workings of his mind would be if the true state were not ω, but some other state ζ. It obviously makes no sense to postulate (P2) or (P3) in such a large world, but these assumptions are needed if knowledge is to be modeled in terms of the possibility partitions of Bayesian decision theory.

6 Conclusion

There is no problem in requiring knowledge to be both justified and true in a small world. This paper argues that the same may be possible in large worlds, but only at the expense of abandoning much of the structure that Bayesian decision theory takes for granted.

[2]Note that possibility in Section 3 is taken to be the failure to get a NO when the justification machine is asked whether something is possible. But this is not the same as getting a YES to the same question.

References

[1] R. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55:1–18, 1987.

[2] K. Binmore. Modeling rational players I. *Economics and Philosophy*, 3:9–55, 1987.

[3] K. Binmore. Debayesing game theory. In B. Skyrms, editor, *Studies in Logic and the Foundations of Game Theory: Proceedings of the Ninth International Congress of Logic, Methodology and the Philosophy of Science*. Kluwer, Dordrecht, 1992.

[4] K. Binmore. Foundations of game theory. In J.-J. Laffont, editor, *Advances in Economic Theory*, Cambridge, 1992. Sixth World Congress of the Econometric Society, Cambridge University Press.

[5] K. Binmore. *Fun and Games*. D. C. Heath, Lexington, MA, 1992.

[6] K. Binmore and H. Shin. Algorithmic knowledge and game theory. In C. Bicchieri and M. Chiara, editors, *Knowledge, Belief and Strategic Interaction*. Cambridge University Press, Cambridge, 1992.

[7] P. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[8] D. Lewis. Elusive knowledge. *Australian Journal of Philosophy*, 74:549–567, 1996.

[9] L. Savage. *The Foundations of Statistics*. Wiley, New York, 1951.