

Knights, knaves and unknowable truths

ROY T. COOK

1. Introduction

The fact that we cannot know, of a particular sentence Φ , that Φ is both true and unknowable (or both true and unknown) is the heart of a number of paradoxes, including the paradox of knowability (see Fitch 1963) and the paradox of the knower (see Kaplan & Montague 1960). In essence, the problem is this: If knowability (or knowledge) distributes across conjunction:

$$K(\Phi \ \& \ \Psi) \Rightarrow K(\Phi) \ \& \ K(\Psi)$$

and knowability¹ (or knowledge) is factive:

$$K(\Phi) \Rightarrow \Phi$$

then the claim that we can know, of some sentence Φ , that Φ is both true and unknowable (or true and unknown) leads to a contradiction:

- | | |
|--------------------------------------|-------------------------------|
| (1) $K(\Phi \ \& \ \neg K(\Phi))$ | Assumption |
| (2) $K(\Phi) \ \& \ K(\neg K(\Phi))$ | (1), distributivity |
| (3) $K(\Phi)$ | (2), $\&$ -elimination |
| (4) $K(\neg K(\Phi))$ | (2), $\&$ -elimination |
| (5) $\neg K(\Phi)$ | (4), factivity |
| (6) \perp | (3), (5), \neg -elimination |

In other words, the following claim (the *Constructive Unknowable Propositions Principle*)² is inconsistent (where the quantifiers range over sentences or propositions):

$$C: (\exists P)K(P \ \& \ \neg K(P))$$

Debate remains open, however, regarding whether we can know that there are unknowable truths in a less constructive way. In other words,

¹ Sometimes knowability, unlike actual knowledge, is not factive. For example, it is plausible that 'I am now thinking this very sentence' is knowable at any time by any sufficiently rational agent, even though it is not necessarily true at those times. Here I ignore such phenomena, however, since these sorts of epistemic puzzles, although interesting and important in their own right, seem orthogonal to the issue at hand.

² In the remainder of this essay I will, for simplicity's sake, concentrate on knowability rather than actual knowledge. The points made below should straightforwardly apply to knowledge as well, however.

there seem to be, as of yet, no decisive arguments³ for or against the coherence of the *Non-constructive Unknowable Propositions Principle*:

$$\mathbf{N}: K(\exists P)(P \ \& \ \neg K(P))$$

The purpose of this note is to present some tentative, yet, I think, extremely suggestive, considerations in favour of the view that there are (or, at least, could be) unknowable truths in the latter, weaker sense.

The trick, of course, is determining how we might come by a warrant for the truth of **N** without that evidence also being a warrant for the truth of **C**. The answer, or at least one answer, is that, although we cannot identify unknowable truths individually (on pain of contradiction), perhaps we can identify classes of sentences which are guaranteed to contain some unknowable truths (where we cannot identify which members of the class are true but unknowable).

To see how this might work, consider the *Disjunctive Constructive Unknowable Propositions Principle*:

$$\mathbf{D}: (\exists P)(\exists Q)K[(P \ \& \ \neg K(P)) \vee (Q \ \& \ \neg K(Q))]$$

If we could find two sentences *P* and *Q* such that (i) we knew that one or the other was true and unknowable, and (ii) we could not tell, of either *P* or *Q* alone, that it was definitely true and unknowable,⁴ our problem would be solved. Given standard assumptions regarding the interaction of the knowability operator and existential quantification, **D** implies **N** (since if there are two propositions such that we know one or the other of them is true but unknowable, then we certainly know that there is some sentence which is true but unknowable). **D** by itself does not imply **C**, however (since there is no single sentence which we know to be true but unknowable – we know at least one of two is, but don't know which).

In what follows we shall actually look at a slightly weaker principle, however. The principle in question is the following *Doubly Disjunctive Constructive Unknowable Propositions Principle*:

$$\mathbf{D2}: (\exists P)(\exists Q)(\exists R)K[(P \ \& \ \neg K(P)) \vee (Q \ \& \ \neg K(Q)) \vee (R \ \& \ \neg K(R))]$$

We should note that **D2**, and similar principles such as **D**, are, strictly speaking, stronger than our actual target **N**. Thus, although **D2** entails **N**, an argument for the incoherence of former would say little about the more general status of the latter. Thus, if the argument presented below

³ Of course, there are plenty of non-decisive arguments.

⁴ If we accept that the knowledge of a disjunction entails knowledge of one or more of the disjuncts, as do some anti-realists, then the distinction at hand collapses, we are back to **C**.

⁵ The reasons for focusing on **D2**, rather than **D**, are technical, and beyond the scope of this note. The interested reader is encouraged to consult Cook (in preparation).

fails, this does not entail the non-existence of unknowable truths. Instead, such a failure merely shows that the present strategy fails, and we do not know (currently) that there are such truths.

In what follows I will first exhibit the basic argument in terms of a puzzle occurring on Raymond Smullyan's celebrated island of knights and knaves (see, e.g. Smullyan (1978)). In §3 I will briefly discuss the argument and outline what an objection would need to achieve. In the appendix I sketch a formal version of the argument in an extension of Peano Arithmetic.

2. *Knights, Knaves, and Knowability*

First discovered by Raymond Smullyan, the island of knights and knaves is the home to two tribes who lend their names to the island. The first tribe, the knights, always tell the truth, while the other, the knaves, always lie.⁶ Otherwise members of each tribe are indistinguishable from each other (in what follows it is assumed that you know these facts about the islanders).

Suppose that you come across three natives of the island, whose names, you discover, are Ava, Brigitte and Dorothy.⁷ After some idle chit-chat, the three island natives suddenly make the following utterances simultaneously:

Ava: What Brigitte is now saying cannot be known to be true.

Brigitte: What Dorothy is now saying cannot be known to be true.

Dorothy: What Ava is now saying cannot be known to be true.

Can we draw any interesting conclusions about Ava, Brigitte, and Dorothy? The answer, it turns out, is 'yes'. First off, we have:

Lemma: At least two of Ava, Brigitte, and Dorothy are knights.

Proof: For *reductio*, assume that no more than one of Ava, Brigitte, and Dorothy is a knight. Then at least two of them are knaves. Without loss of generality, assume that Ava and Brigitte are knaves. So, what Ava said must be false. So, since Ava said 'What Brigitte is now saying cannot be known to be true', what Brigitte said can be known to be true. By factivity, this implies

⁶ It is important, both in what follows and in Smullyan's original puzzles, that neither knights nor knaves are allowed to make utterances that will bring about paradoxes.

⁷ Note that a native of the island can communicate their name to you by uttering, e.g. 'I am a knight if and only if my name is Ava'. More generally, any truth Φ can be communicated by an utterance of the form 'I am a knight if and only if Φ ' without thereby revealing the tribe of the utterer (unless Φ itself contains such information).

that what Brigitte said is true. But Brigitte is a knave, and cannot utter truths. Contradiction. So at least two of Ava, Brigitte and Dorothy are knights.

This lemma leads directly to the following:

Theorem: At least one of the utterances of Ava, Brigitte and Dorothy is true but unknowable.

Proof: By the lemma, we know that at least two of Ava, Brigitte and Dorothy are knights. Assume that Ava and Brigitte are knights. Then, since Brigitte is a knight, Brigitte's utterance is true. Since Ava is a knight, and Ava uttered 'What Brigitte is now saying cannot be known to be true', it follows that Brigitte's true utterance is unknowable. Since the choice of Ava and Brigitte was arbitrary, this completes the proof.

Thus, letting α , β , and δ represent the utterances of Ava, Brigitte, and Dorothy respectively, we have:

$$K[(\alpha \ \& \ \neg K(\alpha)) \vee (\beta \ \& \ \neg K(\beta)) \vee (\delta \ \& \ \neg K(\delta))]$$

which is an instance of **D2**.⁸

3. Conclusion

To sum up, if it is possible that three beings, each of whom is either an unflinching truth-teller or an unflinching liar, could make the utterances described in the previous section, then there are unknowable truths. Put a bit less contentiously, if three normal beings can simultaneously make utterances similar to those of Ava et al. without any paradox, failure of reference, or other pathology, then there are unknowable truths (assuming that classical logic holds and knowability behaves as described). Of course, this might not be possible in the relevant sense, but there seem to be no obvious reasons for thinking so. Thus, anti-realists and others who

⁸ One interesting corollary is that Ava et al. are now much more constrained in terms of their possible future utterances than they were before. For example, prior to their pronouncements, had Ava been a knight, she could have uttered 'I am either a knight or a knave', telling an obvious truth and thereby alerting us to her status as a truth-teller. In addition, if Brigitte were also a knight, she could have uttered the same sentence, with the same results. After their statements (as described above), however, the situation is more complicated: If Ava and Brigitte are both knights, then only Ava can utter 'I am either a knight or a knave' – Brigitte must restrict herself to pronouncements such as 'I am a knight if and only if my name is Brigitte' which provide no information regarding her knight/knave status. Additionally, if all of Ava, Brigitte and Dorothy are knights (a consistent possibility), then this can never be known.

wish to deny the very possibility of unknowable truths must also tell us what goes wrong in situations such as the one described above.

There is a simple objection to all of this, however, that is worth pointing out. Although the utterances of Ava and her friends can be shown to be consistent relative to the principles listed in §1 (factivity and distributivity), if we add the KK principle⁹ then the utterances of Ava et al. are incoherent (given the rules governing inhabitants of the island):¹⁰

Theorem: The KK principle implies the impossibility of any knights and knaves behaving as described.

Proof: Left to the reader (or see Cook (in preparation)).

Thus, the KK principle entails that the situation described above is impossible, and thereby entails denying that the strategy outlined above produces unknowable truths (or, presumably, any truths at all).

The status of the KK principle, however, is far from uncontroversial, and discussion tends to be tied up with other issues (such as the internalism/externalism debate).¹¹ At any rate, there seems to be dialectical space for a view that would deny both the KK thesis and the existence of unknowable truths. If nothing else, the argument sketched above demonstrates that there will be serious problems with such a view.

Appendix

Assume we are working in the language of Peano Arithmetic (PA) plus a predicate ‘K’ (let us call this language L_K). By the diagonalization theorem (see, e.g. Boolos & Jeffries 1989, esp. §15), we can find three sentences α , β , and δ of L_K such that:¹²

⁹ The KK principle states that if something is knowable, then it is knowable that it is knowable, or $K(P) \rightarrow K(K(P))$. Within modal logic this principle is often called 4.

¹⁰ The epistemic principle known as B (i.e. $P \rightarrow K(\neg K(\neg P))$) also renders the utterances of Ava et al. incoherent. This fact is rendered less interesting than the case of KK since there are no independent reasons for thinking that this principle governs knowability or knowledge (other than acceptance of the stronger principle that truth implies knowability, which is exactly what is at issue.)

¹¹ Of course, the question of internalism versus externalism is not completely orthogonal to whether or not there could be unknowable truths, independently of the arguments presented here.

¹² I suggest that *The Epistemic Open Triple* is an appropriate label for this particular construction, paralleling the now well-known (semantic) open pair:

$$\begin{aligned} \Phi &\leftrightarrow \neg T^{\lceil \Psi \rceil} \\ \Psi &\leftrightarrow \neg T^{\lceil \Phi \rceil} \end{aligned}$$

$$\begin{aligned} \text{PA} &\Rightarrow \alpha \leftrightarrow \neg \text{K}(\ulcorner \beta \urcorner) \\ \text{PA} &\Rightarrow \beta \leftrightarrow \neg \text{K}(\ulcorner \delta \urcorner) \\ \text{PA} &\Rightarrow \delta \leftrightarrow \neg \text{K}(\ulcorner \alpha \urcorner) \end{aligned}$$

The theory PA_K results from the addition of the following rules to PA:

$$\begin{aligned} &[\Phi \text{ and } \Psi \text{ are formulae using solely } \alpha, \beta, \text{ and } \delta \text{ as atoms.}] \\ \text{Factivity}_{\alpha\beta\delta}:^{13} & \quad \text{K}(\ulcorner \Phi \urcorner) \Rightarrow \Phi \\ \text{Closure}_{\alpha\beta\delta}: & \quad \text{K}(\ulcorner \Phi \ \& \ \Psi \urcorner) \Rightarrow \text{K}(\ulcorner \Phi \urcorner) \ \& \ \text{K}(\ulcorner \Psi \urcorner) \\ \text{Necessitation}_{\alpha\beta\delta}: & \quad \text{If: } \quad \Rightarrow \Phi \\ & \quad \text{Then: } \quad \Rightarrow \text{K}(\ulcorner \Phi \urcorner) \end{aligned}$$

We obtain the following results:

Theorem 1: PA_K is consistent

Proof: (see Cook (in preparation))

Lemma 1: $\Rightarrow (\alpha \ \& \ \beta) \vee (\alpha \ \& \ \delta) \vee (\beta \ \& \ \delta)$

Proof: $\neg (\alpha \vee \beta) \Rightarrow \neg (\alpha \vee \beta)$
 $\neg (\alpha \vee \beta) \Rightarrow \neg \alpha$
 $\neg (\alpha \vee \beta) \Rightarrow \text{K}(\ulcorner \beta \urcorner)$
 $\neg (\alpha \vee \beta) \Rightarrow \beta$
 $\neg (\alpha \vee \beta) \Rightarrow (\alpha \vee \beta)$
 $\Rightarrow (\alpha \vee \beta)$

Similarly,

$$\begin{aligned} &\Rightarrow (\alpha \vee \delta) \\ &\Rightarrow (\beta \vee \delta) \end{aligned}$$

So,

$$\Rightarrow (\alpha \ \& \ \beta) \vee (\alpha \ \& \ \delta) \vee (\beta \ \& \ \delta)$$

Theorem 2: $\Rightarrow \text{K}(\ulcorner (\alpha \ \& \ \neg \text{K}(\ulcorner \alpha \urcorner)) \vee (\beta \ \& \ \neg \text{K}(\ulcorner \beta \urcorner)) \vee (\delta \ \& \ \neg \text{K}(\ulcorner \delta \urcorner)) \urcorner)$

Proof: $(\alpha \ \& \ \beta) \Rightarrow \alpha$
 $(\alpha \ \& \ \beta) \Rightarrow \neg \text{K}(\ulcorner \beta \urcorner)$
 $(\alpha \ \& \ \beta) \Rightarrow \beta$
 $(\alpha \ \& \ \beta) \Rightarrow \beta \ \& \ \neg \text{K}(\ulcorner \beta \urcorner)$
 $(\alpha \ \& \ \beta) \Rightarrow (\alpha \ \& \ \neg \text{K}(\ulcorner \alpha \urcorner)) \vee (\beta \ \& \ \neg \text{K}(\ulcorner \beta \urcorner)) \vee (\delta \ \& \ \neg \text{K}(\ulcorner \delta \urcorner))$

For an early and insightful discussion of this puzzle, see Roy Sorensen 2003. Sorensen calls this phenomenon the *No-No paradox*, although the more recent nomenclature ‘open pair’ provides for generalization to similar referential loops of any length that is less awkward than ‘No-No-No paradox’, etc.

¹³ $\ulcorner \Phi \urcorner$ is the Gödel number of Φ , and \Rightarrow is the deducibility relation in PA_K .

Similarly,

$$\begin{aligned}(\alpha \& \delta) &\Rightarrow (\alpha \& \neg K(\alpha)) \vee (\beta \& \neg K(\beta)) \vee (\delta \& \neg K(\delta)) \\(\beta \& \delta) &\Rightarrow (\alpha \& \neg K(\alpha)) \vee (\beta \& \neg K(\beta)) \vee (\delta \& \neg K(\delta))\end{aligned}$$

Combining with *Lemma 1*, we have:

$$\Rightarrow (\alpha \& \neg K(\alpha)) \vee (\beta \& \neg K(\beta)) \vee (\delta \& \neg K(\delta))$$

Which, by necessitation, provides the required:

$$\Rightarrow K[(\alpha \& \neg K(\alpha)) \vee (\beta \& \neg K(\beta)) \vee (\delta \& \neg K(\delta))]$$

Theorem 3: $K(\alpha) \rightarrow K(K(\alpha)), K(\beta) \rightarrow K(K(\beta)) \Rightarrow \perp$

Proof: Left as exercise for reader (or see Cook (in preparation)).¹⁴

*Arché: The AHRC Centre for the Philosophy of
Logic, Language, Mathematics, and Mind
University of St Andrews, St Andrews
Fife KY16 9AL, Scotland, UK*

and

*Villanova University
Villanova, PA 19085, USA
Roy.cook@villanova.edu*

References

- Boolos, G. and R. Jeffries. 1989. *Computability and Logic*, 3rd ed. Cambridge: Cambridge University Press.
- Cook, R. In preparation. Epistemic open pairs, epistemic open triples, and unknowability.
- Fitch, F. 1963. A logical analysis of some value concepts. *The Journal of Symbolic Logic* 28: 135–42.
- Kaplan, D. and R. Montague. 1960. A paradox regained. *Notre Dame Journal of Formal Logic* 1: 79–90.
- Sorensen, R. 2003. A definite no-no. In *Liars and Heaps*, ed. JC Beall, 225–29. Oxford: Oxford University Press.
- Smullyan, R. 1978. *What is the Name of this Book?* New York: Prentice Hall.

¹⁴ An early version of this material was presented at Arché: The AHRC Centre for the Philosophy of Logic, Language, Mathematics, and Mind at The University of St. Andrews. I am grateful for the generous environment and helpful feedback always found there.