Epistemic Consistency in Knowledge-Based Systems (extended abstract)

Ron Chrisley
Centre for Cognitive Science,
Sackler Centre for Consciousness Science, and
Department of Informatics
University of Sussex, Falmer, United Kingdom
ronc@sussex.ac.uk

1 Introduction

One common way of conceiving the knowledge-based systems approach to AI is as the attempt to give an artificial agent knowledge that *P* by putting a (typically linguaform) representation that means *P* into an epistemically privileged database (the agent's *knowledge base*). That is, the approach typically assumes, either explicitly or implicitly, that the architecture of a knowledge-based system (including initial knowledge base, rules of inference, and perception/action systems) is such that the following sufficiency principle should be respected:

- **Knowledge Representation Sufficiency Principle (KRS Principle)**: if a sentence that means *P* is in the knowledge base of a KBS, then the KBS knows that *P*.

The KRS Principle is so strong that, although it might be able to be respected by KBSs that deal exclusively with a priori matters (e.g., theorem provers), most if not all empirical KBSs will, at least some of the time, fail to meet it. Nevertheless, it remains an ideal toward which KBS design might be thought to strive.

Accordingly, it is commonly acknowledged that knowledge bases for KBSs should be consistent, since classical rules of inference permit the addition of any sentence to an inconsistent KB. Accordingly, much effort has been spent on devising tractable ways to ensure consistency or otherwise prevent inferential explosion.

2 Propositional epistemic consistency

However, it has not been appreciated that for certain kinds of KBSs, a further constraint, which I call *propositional epistemic consistency*, must be met. To explain this constraint, some notions must be defined:

- An *epistemic* KBS is one that can represent propositions attributing propositional knowledge to subjects (such as that expressed by "Dave knows the mission is a failure").
- An *autoepistemic* KBS is an epistemic KBS that is capable of representing, and therefore of attributing propositional knowledge to, itself (e.g., "HAL knows that Dave knows that the mission is a failure" in the case of the KBS HAL).

All autoepistemic systems (natural or artificial) suffer from epistemic blindspots (Sorensen, 1984):

- A proposition P is an an *epistemic blindspot* for a KBS X if P is consistent, but the proposition that X knows that P is not consistent.

Thus, if an autoepistemic KBS is to respect the the KRS Principle, no epistemic blindspots (for that *KBS*) can appear in its knowledge base.

Despite this, it is of course not logically impossible that a sentence S expressing an epistemic blindspot for a KBS X may end up in X's KB. If this were to happen, it follows that X would not respect the KRS Principle. Worse, the fact that epistemic blindspots are consistent means that this possibility remains *even if X has perfect, ideal methods of normal* consistency *maintenance*. S being in X's KB yields a kind of inconsistency distinct from normal inconsistency (since it can occur even when X's KB, including S, is consistent). Accordingly, X's KB being free of epistemic blindspots for X is a kind of consistency beyond consistency *simpliciter*; this is what I call *propositional epistemic consistency*. To ensure that a *KBS* respects the KRS Principle, then, it is not sufficient to ensure that its KB is consistent in the normal manner; one must also ensure that it is propositionally epistemically consistent.

Ensuring propositional epistemic consistency for a KBS X amounts to taking two precautions:

1. Ensuring that there are no epistemic blindpsots for X in the initial KB;
2. When any sentence S is about to be added to the KB (via inference, perception, etc), checking that S is not an epistemic blindspot for X.

Both steps involve checking that a given sentence is not an epistemic blindspot for a given system X. Beyond checking the consistency of S (and the consistency of S with the current KB), this amounts to checking whether it would be a contradiction to suppose that S is known by X. In turn, this amounts to expressing S in conjunctive normal form, where the first conjunct is the proposition P, and the second conjunct of is of the form $\neg K(x,P)$, where x refers to X.

Unfortunately, this last condition implies that unlike for consistency simpliciter, checking for propositional epistemic consistency cannot proceed purely syntactically. Simple consistency is a matter of what holds in all models, and is therefore an *a priori* matter independent of the state of affairs in the actual world. But whether or not an expression in fact refers to a given individual does depend on the state of affairs in the actual world, and cannot be determined via *a priori* means alone.

In the face of this apparent intractability, and the fact that it derives from a kind of unrestricted self-reference, one might be tempted to reduce propositional epistemic

consistency checking to simple consistency checking in a way parallel to the way Prior proposes for dealing with the paradox of the liar. Prior suggests that we understand each sentence to be implicitly asserting "this sentence is true"(Prior, 1976). This renders such sentences as "This sentence is not true" as straightforwardly false, and thus non-paradoxical. A parallel move would be to suggest that every KBS's KB is implicitly asserting the negation of every epistemic blindspot for that KBS. This would render every epistemic blindspot for that KBS inconsistent with that KBS's KB, allowing it to be excluded via simple consistency maintenance. But this is overkill: epistemic blindspots are not, in general, false. And the ones that are problematic are so because they are true, so having their negations in the KB violates the KRS Principle.

3 Inferential epistemic consistency

There are similar, problematic interactions concerning inference. Consider inference *G*:

1. HAL has made more than two inferences
2. HAL has made fewer than four inferences
3. If someone has made more than two inferences and fewer than four inferences, they have made three inferences
4. Therefore, HAL has made three inferences

On the face of it, *G* is a valid argument; the rules of inference it employs are valid in that they guarantee the truth of the conclusion, given the truth of the premises. And such an analysis is correct (or at least seems so) for the case of you or I putting forward *G*, or making the inference it licenses. But the case of HAL carrying out this inference is another matter entirely. If HAL makes this inference, HAL comes to believe something false, since after the inference is made, HAL believes that HAL has made three inferences, when in fact HAL has made four. HAL's KB would exhibit *inferential epistemic inconsistency*.

On the standard view, one makes an inference by first determining if the premises are true and the transitions from premise to conclusion are valid. If they are, then one should believe the conclusion. Unfortunately, such an approach would license HAL to make inference *G*.

Prompted by these considerations, and taking a more participatory view of inference, I propose that in when one is about to make an inference, in addition to checking for soundness and validity of an inference, one should consider the nearest possible world in which one carries out the inference. Only if the conclusion still follows validly from true premises in that world should one make the inference and believe the conclusion (in this world). On this view, HAL would not be entitled to make the inference in *G*, as its conclusion is false in the nearest possible world in which HAL makes the inference.

Notice that, like the epistemic blindspots considered earlier, the conclusion of *G* that HAL is not entitled to believe is, nevertheless, consistent: possibly true. The conclusion is not, however, a blindspot: the proposition that HAL *knows* the conclusion of *G* is not a contradiction. Nor is it just an inferential variation on an epistemic blindspot.

4 Conclusion

The primary conclusion of the foregoing is that designers of autoepistemic KBSs must supplement consistency checks with epistemic consistency checks of two kinds (propositional and inferential) in order to:

- Respect the KRS Principle that underlies all KBS use;
- Ensure the validity of inferences KBSs make about themselves;
- Ensure consistency of KBS knowledge bases;
- Prevent the introduction of false propositions into KBS knowledge bases.

References

Prior, A. (1976). *Papers in logic and ethics*. Duckworth.

Sorensen, R. (1984). Conditional blindspots and the knowledge squeeze: a solution to the prediction paradox. *Australasian J. Phil.*, *62*, 126-135.