Epistemic blindspot sets: A resolution of Sorensen's strengthened paradox of the surprise examination

Ron Chrisley

June 13, 2006

1 Sorensen's strengthened paradox

A teacher says to Dave: starting next week, I will pay you $1000 if at midnight you have the intention to take an exam on the next day. Furthermore, as long as you carry out your intention, this offer will be renewed the next day, for every day next week.

We need to assume some reasonable facts about Dave to turn this into a paradox:

- He does not like taking exams, but he does like money; he certainly doesn't mind taking an exam, *a fortiori* doesn't mind merely intending to take an exam, if it means he'll get $1000;
- He's not stupid; he won't do something he doesn't like if it doesn't gain him anything;
- He knows he's not stupid (in the sense described above);
- He is *au fait* with the results of the philosophy of action; in particular, he knows: (I1) $\forall x \forall a(K(x, \neg D(x, a)) \rightarrow \neg I(x, a))$

What is the paradox, exactly?

Parallels with Newcomb's Problem.

Assumption of non-ballistic rationality.

2 My solution

Let S be a set of propositions. Let S_X be just the set of propositions that assert, for each member p of S, that X knows that p. Given this, we can say that a set of propositions S constitutes an *epistemic blindspot set for* X iff S is consitent, but S_X is not.

S = {It is raining, but Jim doesn't know it}, then S_{Jim} = {Jim knows that it is raining but Jim doesn't know it}.

S = {If Ralph survived, Ralph is the only one who knows it; Ralph survived}, then S_{Jim} = {Jim knows that Ralph survived; Jim knows that if Ralph survived, Ralph is the only one who knows it}.

It follows from this that if S is an epistemic blindspot set for someone X, then X cannot simultaneously know all the members of S.

Since Dave is not stupid, he will not take the exam on Friday:

(S1) $\neg D(d, F)$.

Since Dave knows he is not stupid, he knows (S1):

(K1) $K(d, \neg D(d, F))$.

Moreover, since Dave is ideally rational, he knows (I1):

$\forall x \forall a(K(x, \neg D(x, a)) \rightarrow \neg I(x, a))$.

But in order to apply this in his reasoning, he has to instantiate it as:

$K(d, \neg D(d, F)) \rightarrow \neg I(d, F)$.

But this is equivalent to:

(S2) $I(d, F) \rightarrow \neg K(d, \neg D(d, F))$.

(I2) $\forall x \forall a(I(x, a) \rightarrow K(x, I(x, a)))$

Now as soon as Dave forms the intention to take the exam on Friday, we have:

(S3) $I(d, F)$

which, together with (I2), implies:

(K3) $K(d, I(d, F)))$

But $S = \{S1, S2, S3\}$ is an epistemic blindspot set for Dave.

It follows that since Dave knows S1 and S3 (cf (K1) and (K3)), he cannot know S2.

He is therefore barred from knowing any conclusions that he draws from S2.

In particular, he cannot conclude that he cannot form an intention to take the exam on Friday. So the reasoning is arrested there, and the paradox is resolved.