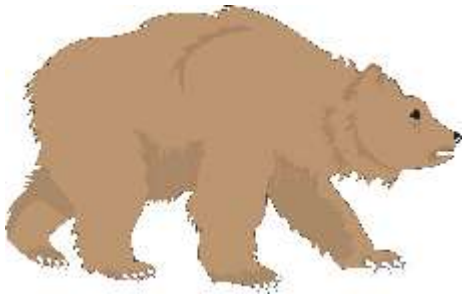
Philosophical Communications, Web Series, No. 32, pp. 132-136.
Dept. of Philosophy, Göteborg University, Sweden
ISSN 1652-0459

Ursus Philosophicus

Essays dedicated to Björn Haglund on his sixtieth birthday



The Magic of Negative Introspection

Staffan Larsson

Abstract

I will show that if combined with the classical analysis of knowledge as (justified) true belief, the axiom of negative introspection has some very counterintuitive consequences. Specifically, if an agent capable of negative introspection has a belief in a proposition which is in fact false, she will also believe that she does not know that the proposition is false. In addition, any agent for which the axiom of negative introspection holds would not feel certain of any false propositions, i.e., she would be incapable of having false strong beliefs.

1. Introduction

The modal logic S5 is frequently used for axiomatising the notion of knowledge, e.g. in Fagin et. al. (1995) and Hoek and Mayer (1989). Among its axioms is one often referred to as "negative introspection", which claims that if one does not know ϕ, then one knows that one does not know ϕ, or formally

$$K\phi \rightarrow K\neg K\phi.$$

I will show that if combined with the classical analysis of knowledge, describing knowledge as (justified) true belief, this axiom has some very counterintuitive consequences and should therefore not be used to characterise knowledge.

2. An intuitive argument

An intuitive argument runs as follows: Suppose I believe ϕ, but ϕ is false. Then, I do not know ϕ, since knowledge implies truth – I cannot know anything that is false. By the axiom of negative introspection I know that I do not know that ϕ, and thus I believe that

I do not know ϕ. This is slightly unsettling, but things get even worse. On a strong notion of belief, “believing that ϕ” can be paraphrased as “being certain that ϕ”. Given this, it seem plausible to say that one cannot at the same time believe (be certain) that ϕ and believe that one does not know ϕ. Presumably, if one felt certain that ϕ, one would also feel quite certain (or at least believe) that one actually knew ϕ. This would imply that any agent for which the axiom of negative introspection holds would not feel certain of any false propositions, i. e., would be incapable of having false strong beliefs. This is the magical property of negative introspection property that we might call *factual faultlessness*.

Note that this is a claim of a completely different character than the well-known problem of logical omniscience. The logic S5 and similar logics imply that agents know all logical truths, which includes all truths of mathematics, and that an agent's knowledge is closed under logical consequence. These consequences can all be blamed on the assumption of ideal rationality; factual faultlessness, however, has nothing to do with rationality. Instead, it appears to require some magical ability to “perceive the world directly”.

The apparent plausibility of the axiom as a characteristic of knowledge comes, I suspect, partly from it's applicability to the closely related concept of belief. Also, people often give the answer “I don’t know”, indicating that they at least know that they don’t know. The problem is that there are at least two ways of not knowing - not believing, or not being correct in ones belief. Negative introspection works fine as long as the cause of lack of knowledge is lack of belief. However, if there is a belief which is false the axiom suggests that there is some magical way that this falsity can affect the beliefs of the believer.

For example, let’s say you know that your car is parked in the street. That is, you feel certain that it is there, you are justified to think so - actually, you left it there just ten minutes ago - and the belief is correct. You also (according to positive introspection) know (at least implicitly) that you know that the car is so parked. Now assume someone steals your car and drives off with it, without you knowing it or having received any information whatsoever that your car was stolen. According the axiom of negative introspection, something will now have happened in your mind - you have acquired a

belief (knowledge, in fact) that you do not know that you car is parked in the street. Thus, the axiom suggests that there is some way that events in the physical world can affect our beliefs without the presence of an act of perception or indeed any causal chain of physical events.

To make clearer how the conclusion of factual faultlessness is derived and what the presupposed axioms are, I will present a formal version of the argument.

3. A formal proof

The following axioms of knowledge are used in Fagin et. al. (1995) as a foundation for an extended exposition of epistemic logic:

(T) $K\phi \rightarrow \phi$

(P) $K\phi \rightarrow KK\phi$

(N) $\neg K\phi \rightarrow K\neg K\phi$

While Fagin et. al. do not formally define the relation between belief and knowledge, the classical analysis of knowledge would at least yield these axioms, claiming that knowledge implies belief and that beliefs are consistent, respectively:

(KB) $K\phi \rightarrow B\phi$

(NB) $B\neg\phi \rightarrow \neg B\phi$

Let's assume that an agent obeying these axioms has a false belief in ϕ, i.e. ϕ is false but believed to be true. We then have the following:

(1)	$\neg\phi$	(assumption)
(2)	$\neg K\phi$	(1, T)
(3)	$K\neg K\phi$	(2, N)
(4)	$B\neg K\phi$	(3, KB)
(5)	$\neg\phi \rightarrow B\neg K\phi$	(1, 4)

Apart from a twist on the logical omniscience problem, claiming that for all false propositions the agent must believe that she does not know them, we can see that if an agent capable of negative introspection has a belief in a proposition which is false, she will also believe that she does not know that the proposition is false. We can now proceed by formalising a notion of strong belief, or (subjective) certainty:

(C) $C\phi \rightarrow BK\phi$

Now, assume that our agent feels certain that ϕ is true, but ϕ is actually, as it happens, false.

(1')	$C\phi$	(assumption)
(2')	$\neg\phi$	(assumption)
(3')	$B\neg K\phi$	(5, 2')
(4')	$\neg BK\phi$	(3', NB)
(5')	$\neg C\phi$	(4', C)
(6')	$\perp$	(1', 5')

This is thus an impossible situation for an agent equipped with the magical powers of negative introspection. If ϕ is false, she will never, no matter what, make the mistake of feeling certain that ϕ is true. If somehow she accidentally happened to come to believe (weakly) that ϕ was true, a bit of introspection would yield a reassuring belief that she did not know ϕ. I feel certain that this is not how knowledge works.

Staffan Larsson
Dept. of Linguistics
Box 200
SE 405 30 Göteborg
Sweden
sl@ling.gu.se

References

R. Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. (1995). *Reasoning about knowledge.* Cambridge, MA and London, U.K.: The MIT Press.

Wiebe van der Hoek and John Jules Meyer (1989). Temporalizing epistemic default logic. *Journal of Logic, Language and Information,* 7(3).