

# On the Incompatibility of Negative Introspection and Knowledge as True Belief

## §1 Introductory Remarks

This paper deals with an incompatibility problem in a branch of epistemic logic. The branch of epistemic logic under consideration here is the 'classical' approach that treats epistemic attitudes like operators in alethic modal logic. This approach was inaugurated by Hintikka's pioneering works (Hintikka 1962), and a first comprehensive state of the art review was provided by Lenzen (1978). This approach has been heavily criticized as its rules and axioms (like logical omniscience and deductive closure) are seen by many as epistemologically and psychologically doubtful (many of these criticisms are discussed in Lenzen 1980). Epistemic logic has thus developed into several other approaches and branches which use more recent logical tools like non-monotonic logics or descriptive logics (cf. the various approaches in Laux/Wansing 1995). Nonetheless the classical approach is still alive, as witnessed by its prominent role in the recent textbook *Epistemic Logic for AI and Computer Science* (Meyer/van der Hoek 2004). Within philosophical logic classic epistemic modal logic often serves as starting point in investigating epistemic attitudes and concepts, as witnessed by the survey papers on epistemic logic in two recent companions to philosophical logic (cf. Goble 2001, Jacquette 2002). Therefore the issue raised in this paper here is still relevant. Even more so as it seems to have gone largely unnoticed in the criticism of classic epistemic modal logic.<sup>1</sup>

---

<sup>1</sup> The only exception is the paper by Larson (2004). I was not aware of Larson's work when writing this paper, luckily a referee for *Erkenntnis* pointed it out to me. In §4 I relate my argument to Larson's paper.

The main argument of this paper here is preceded by two stage setting paragraphs. §2 introduces the introspection principles used in epistemic modal logic, and rehearses some of the psychological criticism. Even if this criticism is not endorsed, negative introspection yields unacceptable consequences when added to the other rules and axioms of epistemic modal logic. §3 argues that negative introspection cannot be algorithmic. Again: even if this is not considered as unacceptable more trouble is to come. The main part of the paper (§4) shows that negative introspection turns out to be incompatible with defining knowledge as true belief. Here "incompatibility" does not mean that the combination of negative introspection, epistemic modal logic and a definition of knowledge as true belief is inconsistent. The combination, however, yields theorems for the epistemic concepts which fail to be compatible with our intuitive concepts of knowledge and conviction in a manner way beyond the problems with logical omniscience and deductive closure.

## **§2 Positive and Negative Introspection in Epistemic Modal Logic**

*Autoepistemic reasoning* is reasoning the inferences of which depend on representing one's own state of belief. A cognitive agent engaged in autoepistemic reasoning draws conclusion from introspective beliefs. Such epistemic beliefs express that the cognitive agent has this and that non-epistemic beliefs. If agent  $a$  has the belief "The cat is on the mat" the introspective belief is "I believe that the cat is on the mat" or – without self-representation – "It is believed that the cat is on the mat". Formally this can be expressed using epistemic modal operators like "B" (for belief) or "K" (for knowledge).

One question may be now, how much access and how reliable access some cognitive agent  $a$  has to its non-epistemic beliefs (typically called 'first order beliefs' as they do not involve epistemic operators). Let  $B$  be the set of the agent's beliefs. An agent with ideal self-access or ideal introspective capacities may fulfil both of

(i) *positive introspection*:  $\alpha \in B \Rightarrow B\alpha \in B$

(ii) *negative introspection*:  $\alpha \notin B \Rightarrow \neg B\alpha \in B$

further on, the ideal agent may also fulfil some version of logical omniscience or deductive closure with respect to its first order and autoepistemic beliefs:

(iii)  $\vdash \alpha \Rightarrow B\alpha \in B$

(iv)  $\vdash (\alpha \supset \gamma), B\alpha \in B \Rightarrow B\gamma \in B$

For human agents this seems way to unrealistic: neither do we believe or know all logical truths, nor are our beliefs closed under logical consequence. The principles of positive and negative introspection can also be expressed as principles of iterating epistemic modal operators:

(v)  $B\alpha \supset BB\alpha$

(vi)  $\neg B\alpha \supset B\neg B\alpha$

One can now recognize that they are epistemic variants of the modal axioms<sup>2</sup> characterising the alethic modal systems S4 and S5:

(vii)  $\Box\alpha \supset \Box\Box\alpha$

(viii)  $\Diamond\alpha \supset \Box\Diamond\alpha$

These are the stronger modal systems. Especially negative introspection seems to require that we believe of *all sentences of the language* that we have no corresponding belief iff we do not have such a belief.

For technical systems (artificial cognitive agents) this might be feasible. If we consider a database, we may say that the facts stored in the database are its first order beliefs. A query

---

<sup>2</sup> Substitute  $\neg\alpha$  for  $\alpha$  in (vi) and remember that  $\Diamond\alpha \equiv \neg\Box\neg\alpha$ .

is a form of introspective access. If the queried fact is stored the positive reply exhibits positive introspection, a negative reply exhibits negative introspection.

Some try to defend the introspective principles by distinguishing between occurring and dispositional belief, or maybe implicit belief. They claim then that we have at least the implicit belief that we believe if we have a occurring belief. All operators, however, have to be read the same way. And then we find a kind of dilemma: With respect to occurring belief it is comprehensible that we also believe that a belief is occurring. Positive introspection seems to correspond to the conscious awareness of an occurring belief. As positive introspection has to apply to any belief, however, we face then an infinity of ever more involved meta-beliefs (just insert 'B $\alpha$ ' for ' $\alpha$ ' any time you like in (v) or (i)). We do not find this infinite hierarchy of meta-beliefs *occurring* in us. We may accept an infinite *disposition* to answer any questions about our (meta-)beliefs with an iterated belief-operator. For dispositional belief positive introspection seems less damaging, but it also seems less compelling. Why should we always have a reliable disposition to a (further) meta-belief corresponding to just any dispositional belief? Our introspective access to occurring beliefs has nothing to do with this capacity. A database fulfils positive introspection because of a reliable look-up in a limited fact storage. Humans may not be built that way. Even positive introspection may thus be too demanding for human agents. For the sake of the argument let us set these worries to the side and see what might happen if we – or some other cognitive agent – had introspection as in the introspection principles.

### **§3 Negative Introspection Cannot Be Algorithmic**

Negative introspection looks even worse than positive introspection, especially when combined with deductive closure: By recognising that you do not believe  $\gamma$ , but believe  $\alpha$ , you will immediately know that  $\gamma$  does not follow from  $\alpha$  (given your other beliefs as well)! As

we also have false beliefs this does not amount to a decision procedure, but if some cognitive agent had *no* contingent beliefs at all, but fulfilled both the closure principles and the introspection principles (i.e. (i) – (iv) above), that agent would constitute some kind of a decision procedure for *any* underlying logic  $\Delta$ , which should give as a pause.

The procedure would be the following: The sentences of a language  $L$  are recursively enumerable (by some Turing machine  $M_1$ ); for good measure the theorems of some undecidable logic  $\Delta$  expressed in  $L$  are recursively enumerable (by some Turing machine  $M_2$ ). Let  $M_1$  provide a sentence  $\alpha$ . Check:  $B\alpha \in B$ ? Either the sentence is believed or it is not the case that it is believed. Even if belief does not obey Excluded Middle (i.e. we may neither believe a sentence nor its negation), *having* a belief is not vague (i.e. obeys Excluded Middle: either we have a belief or we do not). If the sentence is believed, positive introspection tells us so. As, by assumption, the system has no contingent beliefs, but only beliefs delivered by the rule of logical omniscience, we now know that the sentence in question is a theorem. If the sentence is not believed, negative introspection tells us so. Again, as the rule of logical omniscience is the only belief generating rule in the system in question, we can contrapose (the sentence is not believed) and derive that the sentence is not a theorem. Thus the non-theorems are recursively enumerable as well. Any sentence can be decided as to its theoremhood. This or negative introspection alone in combination with the workings of  $M_2$  provides us for any sentence  $\alpha$  with an answer whether in  $\Delta \vdash \alpha$  or  $\not\vdash \alpha$ . This does not provide a decision procedure in the strict sense (and thus no refutation of or contradiction to the undecidability theorems) as the checking procedure certainly is not *algorithmic* – put otherwise: it *cannot* be algorithmic on pains of contradicting undecidability theorems. In so far as negative introspection is the crucial ingredient in this generic decision procedure negative introspection cannot be algorithmic.

Real databases despite their supposed introspective capabilities provide no such problem as they are *finite* and undecidability comes only with infinite domains (here: infinite belief storage spaces). Real databases also only *approximate* logical omniscience stepwise by trying to arrive at a theorem (one compatible with their finite storage space and other limits) by a deduction when prompted for an answer.

#### §4 Negative Introspection and Knowledge as True Belief

Introspection principles are *controversial* in light of human epistemology and human self-access. They are *disastrous* if combined with a strong concept of knowledge. The strong concept of knowledge defines KNOWLEDGE as true conviction:

$$(K_+) \quad K_+\alpha \equiv C\alpha \wedge \alpha$$

This conception has often been criticized as it allows for good luck and happy guesses to count as knowledge. The alternative common conception is to require that only grounded or justified convictions count as knowledge:

$$(K_F) \quad K_F\alpha \equiv C\alpha \wedge \alpha \wedge F\alpha$$

The first difficulty now is to spell out FOUNDATION or JUSTIFICATION (i.e. is it foundationalist or holistic etc.). The second and major difficulty, however, are counter examples concerning cases of having founded convictions and being convinced because of these foundations, but being convinced because of the wrong reasons. Fred is convinced that mail has arrived, because he has seen the post man walking away from his home, and this usually indicates that mail has arrived; indeed mail has arrived, but it was delivered by a substitute post man earlier, the post man Fred knows is on holiday and just came along by accident; one may doubt that

Fred *knows* that mail has arrived, because he is convinced because of the wrong reasons.<sup>3</sup>

Lenzen<sup>4</sup> has highlighted the problem that a proponent of  $(K_F)$  now faces a dilemma: either the procedures and conditions of foundation still allow for counterexamples, or the condition “ $F\alpha$ ” has to be strengthened to a degree that

$$(1) \quad F\alpha \supset \alpha$$

But if this holds,  $(K_F)$  implies  $(K_+)$ , and one may want to see a concept of foundation that strong.

Lenzen, therefore, sees  $(K_+)$  as the natural starting place for epistemic logic. What other principles may now be demanded for the “ $K$ ”-operator?

The definition and our ordinary understanding give the (T) axiom:

$$(T) \quad K_+\alpha \supset \alpha$$

Following the heuristic of looking at alethic modal logics and a prior discussion of the logic of CONVICTION, the (K) axiom and the idealization of logical omniscience (and closure) may be added:

$$(K) \quad K_+(\alpha \supset \gamma) \supset (K_+\alpha \supset K_+\gamma)$$

$$(RK_+) \quad \vdash \alpha \Rightarrow \vdash K_+\alpha$$

For being convinced Lenzen like other epistemic logicians adopts both introspection principles:

$$(C3) \quad C\alpha \supset CC\alpha$$

---

<sup>3</sup> Cases like this were famously raised by Gettier (1963), and have been extensively discussed in analytic epistemology.

<sup>4</sup> Lenzen 1980, pp. 58-61. Conviction is understood here, following Lenzen, as strong belief (i.e. one is convinced that  $\alpha$  if  $\neg\alpha$  does not seem possible to one). One may argue whether this requires too much for a definition of knowledge, but nothing important depends on changing  $(K_+)$  to:  $K_+\alpha \equiv B\alpha \wedge \alpha$ . The introspection principles have been assumed for both belief and conviction.

$$(C4) \quad \neg C\alpha \supset C\neg C\alpha$$

Given (C3) and the definition ( $K_+$ ) we get<sup>5</sup> the (S4) axiom, positive introspection, for strong knowledge:

$$(S4) \quad K_+\alpha \supset K_+K_+\alpha$$

What about the (S5) axiom, negative introspection? The formula would be

$$(S5^*) \quad \neg K_+\alpha \supset K_+\neg K_+\alpha$$

Translating this using the definition ( $K_+$ ) we get:

$$(S5^{*'}) \quad \neg(C\alpha \wedge \alpha) \supset \neg(C\alpha \wedge \alpha) \wedge C(\neg(C\alpha \wedge \alpha))$$

This is unacceptable. The usual way in which we fail to know is being convinced that  $\alpha$ , but  $\alpha$  not being the case:  $C\alpha \wedge \neg\alpha$ . Assume this to be the case. Then the antecedent is true thus we have the consequent. Again the first conjunct of the consequent is unproblematically true. But we have the second conjunct as well. By propositional logic “ $\neg(C\alpha \wedge \alpha)$ ” is equivalent to “ $\alpha \supset \neg C\alpha$ ”. Now by the (K) axiom for “C” we have:

$$(2) \quad C(\alpha \supset \neg C\alpha) \supset (C\alpha \supset C\neg C\alpha)$$

Thus in our assumed case, in which we have  $C\alpha$ , we can detach to get:

$$(3) \quad C\alpha \wedge C\neg C\alpha$$

In combination with (C3) we finally arrive at a contradictory conviction:

$$(4) \quad CC\alpha \wedge C\neg C\alpha \quad \text{or} \quad C(C\alpha \wedge \neg C\alpha)$$

Thus (S5\*) has to be rejected. Only because we have a wrong conviction we certainly have not a contradictory conviction,  $C\perp$ .<sup>6</sup>

---

<sup>5</sup> “ $K_+\alpha$ ” is “ $C\alpha \wedge \alpha$ ” this implies, by (C3), “ $CC\alpha$ ” and using this and the (K) axiom for “C” we get:  $C(C\alpha \wedge \alpha) \wedge (C\alpha \wedge \alpha)$ , i.e.  $K_+K_+\alpha$ .

Lenzen<sup>7</sup> now is not content to have **S4** as the logic of  $K_+$ . He introduces a weakened version of negative introspection, which corresponds to the characteristic axiom of the modal logic **S4.4**.

$$(S4.4) \quad \alpha \supset (\neg K_+ \neg K_+ \alpha \supset K_+ \alpha)$$

Or equivalently:

$$(S4.4') \quad \alpha \supset (\neg K_+ \alpha \supset K_+ \neg K_+ \alpha)$$

This second version shows that we have a weakened form of negative introspection here, one concerning only obtaining facts  $\alpha$ . If we read  $\alpha$  as expressing that  $\alpha$  is a fact (more precisely, that the referent of  $\alpha$  is an obtaining state of affairs), and read  $\neg\alpha$  as expressing that  $\alpha$  is not a fact (more precisely, that  $\alpha$  does not refer to an obtaining state of affairs), or the negative fact that  $\alpha$ 's content is not given, then we can re-formulate (S5\*) as negative introspection for both types of facts:

$$(S5^{**}) \quad \alpha \vee \neg\alpha \supset (\neg K_+ \alpha \supset K_+ \neg K_+ \alpha)$$

We have rejected this principle that does not distinguish obtaining from not obtaining states of affairs, or true or false  $\alpha$ , why should we accept (S4.4') then?

In fact, (S4.4') is as unacceptable as (S5\*\*). To see this assume  $\alpha$  to be true because it refers to some obtaining fact for which you are convinced of the opposite (say the fact of the

---

<sup>6</sup> This argument is also recognized by Larson (2004), who calls it the 'magic of negative introspection' as the presence of negative introspection magically excludes ever being convinced of something which is false, even enforcing 'by magic' the introspective (negative) belief that one does not know (any longer)  $\alpha$  once facts have changed in the world from  $\alpha$  to  $\neg\alpha$ . But this very argument has, of course, earlier been recognized by Lenzen (1980: 61-62) as a champion of negative introspection. The argument presented in this paper here is stronger than Larson's as it is directed at a weaker combination of knowledge principles and negative introspection (the combination Lenzen in fact endorses). Larson's first 'magic of negative introspection' is a variation of the argument above which also has to use the simple (i.e. stronger) version of negative introspection in combination with the other principles of epistemic modal logic.

<sup>7</sup> Lenzen, 1980, pp. 62-65; cf. also Lenzen 1979.

amount of the grains of sand on Mars being even, while you by accident or your queer astronomical methods are convinced that that amount is odd). Now by  $\alpha$  being true we can detach the consequent of (S4.4'). As your conviction is contrary to the facts you do not know  $\alpha$ , i.e.  $\neg K_+\alpha$ . Thus we can detach again and get:  $K_+\neg K_+\alpha$ . You simply have the knowledge that you do not know  $\alpha$ . As you have the conviction  $C\neg\alpha$ , by positive introspection you know that you have that conviction:  $CC\neg\alpha$ . If you ask yourself now "Why do I not know  $\alpha$ " the obvious answer available to you by introspection is "Because I have a contrary conviction". Having understood this much you simply negate your conviction and come to be convinced that  $\alpha$ ,  $C\alpha$ , which again means,  $\alpha$  being the case, that now you come to strongly know  $\alpha$ ! By logic and the obtaining of facts alone one thus comes to revise any old wrong conviction that one has! How could anybody ever have a wrong conviction then anyway – at least have one for an extended period? Therefore, (S4.4) must be rejected. The logic even of  $K_+$  has to be weaker than **S4.4**.

And if that is not bad enough, more trouble is to come.  $K_+\alpha$  obviously implies  $\alpha \wedge \neg K_+\neg K_+\alpha$ .<sup>8</sup> Combining this with (S4.4) gives us:

$$(5) \quad K_+\alpha \equiv \alpha \wedge \neg K_+\neg K_+\alpha$$

If we now write the definition ( $K_+$ ) beneath this

$$(K_+) \quad K_+\alpha \equiv \alpha \wedge C\alpha$$

we immediately see that in the logic of strong knowledge we have the theorem:

$$(6) \quad C\alpha \equiv \neg K_+\neg K_+\alpha$$

And this again is just bizarre: Assume  $\alpha$  to refer to some state of affairs you have never thought of, maybe because it is way beyond your human knowledge. As you have never

---

<sup>8</sup> As " $K_+\alpha$ " implies by (S4) " $K_+K_+\alpha$ " and thus " $K_+\neg K_+\alpha$ " would give a contradiction by the (K) axiom.

thought about  $\alpha$  you certainly do not know  $\alpha$ , i.e.  $\neg K_+ \alpha$ . And as you have never thought about  $\alpha$  you also do not know that you are ignorant or wrong about  $\alpha$ , i.e.  $\neg K_+ \neg K_+ \alpha$ . But then, (6) tells us, you are convinced that  $\alpha$  is true! For everything beyond your ken or interests epistemic **S4.4** for  $K_+$  commits you to a corresponding conviction!

(S4.4) for  $K_+$  thus should be rejected. The moral of our considerations above is: The concept of strong knowledge,  $K_+$ , is not compatible with either unconditioned or conditioned principles of negative introspection for knowledge (i.e. the operator “ $K_+$ ”).

Note that a definition of knowledge containing a justification or foundation clause/requirement as in  $(K_F)$  does not fall prey to the problems presented here so easily. In both problem cases considered we may be very reluctant to see any justification or foundation present – whatever the justification or foundation requirement may finally come down to. Then, however, the critical detachment is not available.

The logic of conviction incorporates a principle of negative introspection, (C4). Apart from the usual criticisms of negative introspection it does not seem to spell problems in the logic of conviction as does negative introspection in the logic of strong knowledge. But this appearance may be deceiving! The logic of conviction **C**, consisting of (C3), (C4), (C2) [the (K) axiom for the operator “C”], a necessitation rule (RC) for the operator “C” [ $\vdash \alpha \Rightarrow \vdash C\alpha$ ], and a consistency principle corresponding to the modal axiom (D):

$$(C1) \quad C\alpha \supset \neg C\neg\alpha$$

if *combined* with the conception of strong knowledge, i.e.  $(K_+)$ , *entails* the other **S4.4** axioms for “ $K_+$ ” as theorems! As also **S4.4** for “ $K_+$ ” in combination with (6) as a definition of “C” entails the logic of conviction, the two logics are provably equivalent.

The direction of the logic of conviction entailing **S4.4** for  $K_+$  is devastating.<sup>9</sup> Since we have rejected (S4.4) we have to reject some of the axioms of the logic of conviction **C** if we want to keep the strong concept of knowledge!

Which of the axioms of **C** have to be given up? Let us look at the problematic direction of the equivalence proof: Axiom (T) follows immediately from  $(K_+)$ , added to **C**.  $(RK_+)$  follows immediately from (RC), the (K) axiom for “ $K_+$ ” follows from (C2) and propositional logic. (S4) follows from (C2) and (C3) as we have argued already above. The crucial part of the proof consists in proving (S4.4).

This proof<sup>10</sup> depends, of course, on (C4), *negative introspection*. Therefore, if we cannot accept (S4.4) and (6), even if we otherwise want to assume a strong conception of knowledge as true conviction, we have to reject negative introspection for conviction *as well*, i.e. give up (C4).

As the other axioms of the strong conception of knowledge can be derived without it we lose only the incriminated axiom (S4.4). If (C4) does not hold, and we drop the accessibility condition that the belief worlds used in modelling someone’s beliefs exhibit a Euclidian structure, we can have “ $\neg C\alpha \wedge \neg C\neg C\alpha$ ” and thus can invalidate (S4.4). As the modal logic

---

<sup>9</sup> Therefore we look only at the proof of this direction of the equivalence.

<sup>10</sup> *Proof:*

|         |   |   |
|---------|---|---|
| 1.<1>   | $\alpha$  | Assumption                                |
| 2.<2>   | $\neg K_+\alpha$  | Assumption                                |
| 3.<2>   | $\neg C\alpha \vee \neg\alpha$                              | $(K_+)$ , 2                               |
| 4.<1,2> | $\neg C\alpha$  | $(\vee E)$ , 1, 3                         |
| 5.<1,2> | $C\neg C\alpha$   | (C4), 4 [ <i>negative introspection</i> ] |
| 6.<1,2> | $C(\neg C\alpha \vee \neg\alpha)$                           | (RC),5                                    |
| 7.<1,2> | $C(\neg K_+\alpha)$   | $(K_+)$ , 6                               |
| 8.<1,2> | $K_+\neg K_+\alpha$   | $(K_+)$ , 2, 6                            |
| 9.<>    | $\alpha \supset (\neg K_+\alpha \supset K_+\neg K_+\alpha)$ | $(\supset I)$ , <u>1</u> , <u>2</u> , 8 ■ |

consisting of  $\mathbf{PC} \cup \{(C1), (C2), (C3), (RC)\}$  – or respectively  $\mathbf{PC} \cup \{(D),(K),(S4),(N)\}$  – is sound, we are assured that there is no proof for (S4.4) in it.<sup>11</sup>

Negative introspection thus is not only a very strong assumption implausible for human reasoners. Negative introspection can not be accepted for a strong conception of knowledge. And because of that it can also not be accepted for the logic of conviction if we stick to the strong conception of knowledge.

---

<sup>11</sup> The system can also be shown to be complete with respect to serial, transitive frames. It is sometimes used also in Deontic Logic, then called “**OS4**”, as the operator “O” (for “it ought to be the case that”) does not obey (T) as well; cf. Åqvist 1987.

## References

- Åqvist, Lennart (1987). Introduction to Deontic Logic and the Theory of Normative Systems. Neapels.
- Gettier, E. (1963). "Is Justified True Belief Knowledge?", *Analysis*, 23, pp.121-23.
- Goble, Lou (2001) (Ed.). *The Blackwell Guide to Philosophical Logic*. Oxford.
- Hintikka, Jaakko (1962). *Knowledge and Belief*. Ithaca.
- Jacquette, Dale (2002) (Ed.). *A Companion to Philosophical Logic*. Oxford.
- Larson, Steffan (2004). "The Magic of Negative Introspection", in: *Ursus Philosophicus*. Essays dedicated to Björn Haglund on his sixtieth birthday. Göteborg.
- Laux, Armin/Wansing, Heinrich (1995) (Eds.). *Knowledge and Belief in Philosophy and Artificial Intelligence*. Berlin.
- Lenzen, Wolfgang (1978). *Recent Work in Epistemic Logic*. Special Issue of Acta Philosophica Fennica, Vol. 30 (1).
- (1979). "Epistemologische Betrachtungen zu [S4, S5]", *Erkenntnis*, 14, pp.33-56.
- (1980). *Glauben, Wissen und Wahrscheinlichkeit*. System der epistemischen Logik. Wien/New York.
- Meyer, J.-J. Ch./van der Hoek, W. (2004). *Epistemic Logic for AI and Computer Science*. Cambridge.