Negative Introspection Is Mysterious

Abstract. The paper provides a short argument that negative introspection cannot be algorithmic. This result with respect to a principle of belief fits to what we know about provability principles.

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).

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*: $\alpha \in B \Rightarrow \mathrm{B}\alpha \in B$

ii) *negative introspection*: $\alpha \notin B \Rightarrow \neg\mathrm{B}\alpha \in B$

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) $\vdash \alpha \Rightarrow \mathrm{B}\alpha \in B$

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

The principles of positive and negative introspection can also be expressed as principles of iterating epistemic modal operators[1]:

v) *positive introspection*: $B\alpha \supset BB\alpha$

vi) *negative introspection*: $\neg B\alpha \supset B\neg B\alpha$

One can now recognize that they are epistemic variants of the modal axioms characterising the alethic modal systems **S4** and **S5**:

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

viii) $\Diamond\alpha \supset \Box\Diamond\alpha$ [equivalent to $\neg\Box\alpha \supset \Box\neg\Box\alpha$ which looks like (vi)]

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) in contrast to human beings 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 is a form of introspective access. If the queried fact is stored the positive reply exhibits positive introspection, a negative reply exhibits negative introspection.

But the databases we know are only finite fact storage, anyway, so the workings of the

[1] The branch of epistemic logic expressing itself in this way 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. 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.

introspection principles seem unsuspicious. In the human case, where we tend to think of the mind as unbounded or at least as a capacity to infinitely many beliefs, the introspection principle are more controversial. Negative introspection looks even worse than positive introspection, especially when combined with deductive closure.

Suppose introspection *and* closure: by recognising that you do not believe γ, but believe α, you will *immediately know* that γ does not follow from α (given your other beliefs as well)! As we ordinary humans 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 Δ, 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 Δ expressed in *L* are recursively enumerable (by some Turing machine M_2). Let M_1 provide a sentence α. 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 α with an answer whether in $\Delta \vdash \alpha$ or $\nvdash \alpha$. In case the logic Δ contains *Modus Ponens* the closure principles have no more import than the theorems being recursively enumerable. The blame thus rests with negative not with positive introspection.

This does not provide a decision procedure in the strict sense (and thus no refutation of or contradiction to the well-known 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*.

This puts doubt on the mere existence of negative introspection as a cognitive capacity. Once we endorse even mild or vague versions of functionalism we suppose that our cognitive and especially our logical/linguistic capacities are program-like. As we have just seen negative introspection cannot be of that kind. A would-be logical capacity of negative introspection is highly mysterious. We should rather forsake its assumption. The argument given above shows that our 'intuitive' complaints against negative introspection can be vindicated by a proper strong argument.

This result corresponds to well-known theorems in provability theory.

If one takes the operator "B" to express provability, then positive introspection is the claim that if α is provable (α is a theorem) then it is provable that α is provable (it is a theorem that α is provable). The provability statement is a so-called 'Σ_1-statement' (expressing the existential claim that there is a proof of α). Any formal system extending the basic arithmetic **Q** is Σ_1-complete. In such a system provability is not only 'expressible' (by the operator or predicate), but is 'semi-representable', which means that that if it is *true* that α is provable it can be *proven* that α is provable. Thus this system fulfils positive introspection. Modal logics of provability therefore incorporate the principle of positive introspection (i.e. the modal axiom characteristic of the system S4).

Negative introspection, on the other hand, is a Π_1-statement (expressing in the context of provability that *all* proofs are such that they do not prove α). Negative introspection thereby exhibits higher complexity than positive introspection. By *Gödel's Second Incompleteness Theorem* negative introspection (for provability) is not only not valid, but is always false, i.e. provability is *at most* semi-representable. If negative introspection had a true instance (for provability) that meant that for some unprovable α it is provable that α is not provable. But proving in a formal system for some formula of the system that it is unprovable in the system means proving the system's own consistency in the system, which is the very thing excluded by *Gödel's Second Incompleteness Theorem*. Modal logics of provability thus exclude the modal axiom characteristic of S5.

To sum up: Negative introspection is unacceptable across the board.

References

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.

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).

Meyer, J.-J. Ch./van der Hoek, W. (2004). *Epistemic Logic for AI and Computer Science*. Cambridge.

Manuel Bremer