< Published in *Analysis* on September 23, 2017. This is the penultimate draft. Please cite the published version. >

Semantic externalism without thought experiments

JUHANI YLI-VAKKURI

1. Semantic externalism (henceforth: externalism) is the thesis that the contents of intentional states (such as beliefs) and speech acts (such as assertions) are not determined by the way the subjects of those states or acts are internally.[1] Externalism is a widely accepted but not entirely uncontroversial thesis. Among widely accepted but not entirely uncontroversial theses in philosophy, externalism is notable for owing the assent it commands almost entirely to thought experiments—especially to Putnam's (1975) 'Twin Earth' thought experiment. The latter is arguably—along with Gettier's (1963) refutation of the justified true belief (JTB) analysis of knowledge—one of the two most influential thought experiments in the history of analytic philosophy.

For this reason, like the view that knowledge is not JTB, one might treat externalism as a test case: is thought experimentation *required* for establishing the view, or are there some widely accepted general principles from which the view can be derived purely by deduction? Recently, Williamson (2013, 2015) has argued that the latter is the case for the view that knowledge is not JTB: the existence of counterexamples to the JTB analysis of knowledge can be shown using certain natural models for epistemic logic. This paper argues for a similar conclusion about externalism: it is a deductive consequence of a pair of widely accepted general principles whose relevance to the issue has so far gone unnoticed.

To avoid any misunderstandings, I should say something about my motives for engaging in this intellectual exercise. A good deal of recent metaphilosophy exhibits hostility towards thought experimentation, as well as towards certain less clearly demarcated practices in the vicinity, variously described as 'appealing to intuitions', 'using intuitions as evidence', making 'judgments about cases', and the like. This paper has no such agenda. I believe that thought experiments can be used to acquire knowledge of important philosophical truths, and that various important philosophical

[1] This is, of course, not the only thesis that the label 'semantic externalism' has been used for. For example, Putnam, who does not use the term in his (1975), is often called a 'semantic externalist', but he argues that the 'meaning' of an utterance of a linguistic expression is not determined by the way the utterer is internally, and it is unclear whether by 'meaning' he means content—which can be specified by 'that'-clauses—or something like 'character' in Kaplan's (1989) sense—which cannot.

truths have, in fact, been discovered by thought experimentation,[2] with externalism itself being a case in point.

However, when thought experiments are dispensable, and they can be replaced by a straightforward deductive argument from premises widely accepted on both sides of a debate, this is well worth knowing because the deductive argument can be more dialectically effective. If one can deduce a claim one's opponents reject from claims they accept, and the validity of the deduction is not in dispute, one is done. The opponents are forced to change their minds about something. If the premises are well chosen—in particular, if they are widely treated as fairly obvious by both sides—it will be especially difficult for the opponents to maintain their opposition. In contrast, a thought experiment scenario is inevitably underspecified in various ways while being overspecified in others—a combination that gives rise to the familiar problem of 'deviant realizations'[3] and tempts the opponents of the thesis supported by the thought experiment to respond by studying (sometimes irrelevant) details of the scenario, and the defenders of the thesis to respond by proposing variants of the original scenario. Diminishing returns can be expected when this process is iterated.

A related point: I take it that everyone agrees that excessive reliance on thought experiments in a particular philosophical literature can have deleterious effects. One risk is that philosophers will, while focusing on the details of thought experiment scenarios, fail to notice certain structural principles that, if explicitly articulated, could significantly move the dialectic forward. Whether, and to what extent, the vast literature inspired by Putnam (1975) is at such an impasse is debatable. But it is clear that that literature has largely focused on the details of Putnam's thought experiment scenario, on coming up with and considering variants of it ('dry earth', 'slow switching', etc.), and on considering various issues tangentially implicated in that scenario, such as the semantics and metasemantics of natural kind terms. As this paper will show, there is no need at all to think through any thought experiment scenarios, so *a fortiori* not ones involving natural kind terms, to appreciate the appeal of externalism.

2. For simplicity's sake, I will focus on the case of belief. Everything I will say about belief could also be said, *mutatis mutandis*, about other propositional attitudes and about speech acts.[4] Externalism will now be understood as the thesis that the content of a belief is not determined by the way the subject of the belief is internally. Or, more precisely, since a subject may have several beliefs not all of which have the same content: the content of a belief is not determined by the way the subject of the

[2] See Williamson (2007: ch. 6) for a defense of this view.

[3] See Williamson (2007: ch. 6) for discussion. The term 'deviant realization' is due to Malmgren (2011). The problem is less pressing when thought experiments are reconstructed, following Williamson, using counterfactual conditionals, but the antecedents of counterfactuals may also be realized in unintended ways.

[4] In some cases the required adjustments will be less obvious than in others. For example, in the case of non-truth-evaluable attitudes like desire, we cannot speak of the attitude having a truth value, as I do in §3. But we can relate desires to truth in a way that will allow the argument of §3 to go through. Necessarily, a desire is *satisfied* iff its content is true. Thus, if we interpret the variables of §3 as ranging over desires, we may interpret *v* as, say, a function that assigns 1 to a desire if it is satisfied and otherwise assigns 0 to it, and assigns 1 to a content if the content is true and otherwise assigns 0 to it.

belief is internally together with the way in which the belief is related to the way the subject is internally.

Beliefs here must be thought of as *tokens* rather than *types*. In this sense, no two subjects share a belief. While there is a natural sense in which you and I share a belief when we both believe that the sky is blue, there is an equally natural sense in which we don't share any beliefs. In this latter sense we may, for example, truly say that my belief was formed by a certain method while yours was not. For this to be true, our beliefs must be distinct, even though they have the same content: that the sky is blue.

The *content* of a belief is what is believed: the content of one's belief that *p* is that *p*. I will make no substantive assumptions about the kinds of entities the contents of belief are, other than that they can be specified by 'that'-clauses: 'that the sky is blue' specifies the content that the sky is blue, 'that 1 + 1 = 2' specifies the content that 1 + 1 = 2, and so on.

Let us now make the thesis of externalism (as applied to beliefs) a bit more precise. Above I glossed it as the thesis that the content of a belief is not determined by the way the subject of the belief is internally and the way in which the belief is related to the way the subject is internally. We can avoid various complications by adopting (following much of the literature) the ideology of *narrowness*, *duplication* and *correspondence*. Things that are internally the same are said to be *duplicates*. Whenever *S* and *S′* are duplicates, there is an intuitive sense in which each part of *S* *corresponds* to a part of *S′*. If *S* and *S′* are normal human subjects, then *S*'s head corresponds to *S′*'s head, *S*'s heart corresponds to *S′*'s heart, and so on. In the same intuitive sense, each belief of one of a pair of duplicate subjects corresponds to a belief of the other. For example, if your belief that snow is white is the only belief you formed during the 25,000,000th time you exhaled, and there is a duplicate of you, then your duplicate's corresponding belief is the only belief he or she formed during the 25,000,000th time he or she exhaled. The relevant notion of correspondence can be made precise in various ways,[5] but for present purposes the intuitive notion will suffice, because nothing in this paper turns on whether some particular actual or possible belief of one of a pair of duplicate subjects corresponds to a belief of the other. Finally, let us say that a property *P* of beliefs is *narrow* iff, necessarily, any corresponding beliefs of duplicate subjects either both have *P* or both lack *P*; otherwise *P* is *broad*. Saying that a property of beliefs is narrow in this sense is one way of making (relatively) precise the idea that whether a belief has that property is determined by the way the subject of the belief is internally together with the way the belief relates to the way the subject is internally.

Clearly not all properties of beliefs are narrow. Clearly not even all semantic properties of beliefs are narrow: truth is a paradigmatic broad semantic property. Content is a disputed case. Externalism can now be precisified as the thesis that

[5] See Yli-Vakkuri and Hawthorne (forthcoming: ch. 1, sec. 3).

content is a broad property of beliefs.[6] I will call the negation of externalism *internalism*.[7]

This way of understanding 'determination' in the initial, rough statement of externalism in §1 is in line with much of the literature, including the *ur*-text (Putnam 1975). Here narrowness is, following Putnam, understood as a species of weak supervenience on the internal.[8] The *A*-properties weakly supervene on the *B*-properties iff it is necessary that any things that have the same *B*-properties have the same *A*-properties. (In the case at hand, having the same *B*-properties amounts to being corresponding beliefs of duplicate agents.[9]) 'Determination' could also be interpreted as strong supervenience, but since a refutation of a weak supervenience thesis is also a refutation of the corresponding strong supervenience thesis, it suffices to focus on weak supervenience here.

3. The argument will be formalized in a language of first-order modal logic with function symbols, identity, and the necessity operator $\Box$, which can be interpreted as expressing metaphysical necessity or any other species of objective necessity (in the sense of Williamson 2017), such as nomological necessity. The formalization of an argument as simple as the one that follows may strike some readers as overkill, but it is helpful in that it enables us to see just which logical principles the argument relies on. It will turn out that the argument will go through in an extremely weak logic, so its validity will presumably not be disputed.

We interpret the formal language so that the variables range over *beliefs*, the two-place predicate C expresses the relation of *being corresponding beliefs of duplicate subjects*, and the function symbols c and v express, respectively, *content* and *truth value*. That is to say, $c(t)$ refers to the content of whatever t refers to and $v(t)$ refers to its truth value.

The formalization of internalism, then, is:

NARROW_C: $\quad \Box \forall x \forall y \, (C(x, y) \rightarrow c(x) = c(y))$

---

[6] This involves some fudging. Content is not a property of beliefs, just as truth value is not (although truth is). Content and truth value are both functions from beliefs to other things. Functions from beliefs to other things are narrow in a derivative sense, which is the sense I use in the text: see Yli-Vakkuri and Hawthorne (forthcoming: ch. 1, sec. 3) for the details.

[7] Who are the internalists? Lewis (1979), Segal (2000), and Farkas (2006, 2008) are perhaps the clearest cases. It is also natural to interpret Searle's (1983: ch. 8, sec. I) claim that 'Intentional content' is 'in the head' as entailing internalism.

[8] Putnam's famous thought experiment is presented as a counterexample to a combination of two supervenience theses, both of which are expressed using 'determine', which is not further explained. It is nevertheless clear that Putnam is using 'determine' to express weak supervenience rather than strong, because one of the two theses is that 'the meaning of a term … determines its extension' (Putnam 1975: 136). Obviously extension at most weakly supervenes on meaning. It is not the case, for example, that, necessarily, if the number of baboons is different from the actual number of baboons, then the meaning of 'baboon' is different from the actual meaning of 'baboon'. There could have been one baboon fewer or more than there actually is without any difference in the meaning of 'baboon', but the strong supervenience of extension on meaning is inconsistent with this fact.

[9] The claim that a certain property is narrow in the present sense is, in fact, equivalent to a weak supervenience thesis of the standard form, but the story of what the *B*-properties are in this case is too complicated to be told here. See Yli-Vakkuri and Hawthorne (forthcoming: ch. 1, sec. 3).

NARROW_C says that, necessarily, any corresponding beliefs of duplicate subjects have the same content, i.e., that content is narrow.

The argument against NARROW_C has just two premises. The first premise is so trifling that it already figured as an example in the stage-setting in §2: it is the assumption that truth is not narrow. This is formalized as:

BROAD_T: $\neg \Box \forall x \forall y (C(x, y) \rightarrow v(x) = v(y))$

BROAD_T says that it is not necessary that all corresponding beliefs of duplicate subjects have the same truth value. Since this assumption passed without comment in §2—truth being a paradigmatically broad semantic property—it will pass without comment here too.

The second premise requires some comment, although it, too, is presumably common ground between externalists and their opponents:

TRANSPARENCY: $\Box \forall x\, v(x) = v(c(x))$

TRANSPARENCY says that, necessarily, the truth value of a belief is the same as the truth value of its content. I don't think this principle is particularly in need of an argument, but here is one anyway:

Suppose for a contradiction that TRANSPARENCY is false. Then it is possible that there is a subject *S* with a belief with a certain content, that *p*, but *S*'s belief that *p* and (the content) that *p* differ in truth value. This can only come about in two ways. The first way is this: *S*'s belief that *p* is true but it is not true that *p* (i.e., that *p* is not true). This is equivalent to: *S*'s belief that *p* is true, and it is not the case that *p*, which is impossible. The second way: *S*'s belief that *p* is not true but it is true that *p* (i.e., that *p* is true). This is equivalent to: *S*'s belief that *p* is not true, and *p*, which is impossible. Both are impossible, so it is not possible for there to a belief that differs in truth value from its content.

It is straightforward to show that the set {TRANSPARENCY, BROAD_T, NARROW_C} is inconsistent, while {TRANSPARENCY, BROAD_T} is consistent, even in an extremely weak system of quantified modal logic that results from combining the weakest normal modal logic **K** with standard first-order logic (FOL). {BROAD_T, TRANSPARENCY, NARROW_C} is inconsistent, for example, in the extraordinarily weak system that we get by taking any standard axiomatization of FOL and adding to it as axioms (i) all instances of the **K** axiom schema $\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$, (ii) all truth-functional tautologies, and (iii) adding the 'necessitation' rule $\phi/\Box\phi$, where ϕ is a closed theorem of FOL (i.e., an FOL theorem with no free variables).[10] Thus there is

---

[10] An Associate Editor and an anonymous referee suggested that I include a proof, so I will. I will adopt the following abbreviations.

A = $\forall x \forall y (C(x, y) \rightarrow c(x)=c(y))$

B = $\forall x\, v(x) = v(c(x))$

C = $\forall x \forall y (C(x, y) \rightarrow v(x) = v(y))$

And I will use '$\vdash$' to express derivability in the system described above.

First note that $(A \wedge B) \rightarrow C$ is a closed FOL theorem. By (iii), then:

(1) $\vdash \Box((A \wedge B) \rightarrow C)$.

By (i), we also have each of:

a derivation of the negation of NARROW_C from TRANSPARENCY and BROAD_T in the system. But NARROW_C expresses internalism, and the negation of internalism is externalism, so there is a derivation of externalism from TRANSPARENCY and BROAD_T in the system. Since TRANSPARENCY and BROAD_T are true, and derivability in the system preserves truth, externalism is true.

4. The argument of §3 is not, of course, psychologically impossible to resist—no philosophical argument is. A sufficiently dedicated internalist will find a way to resist it. Here I will quickly examine two strategies for doing so and indicate why I find them unpromising. (There are, of course, indefinitely many other ways to resist the argument, but the two ways I will discuss are the only ones I have encountered in discussions with internalists and philosophers willing to play devil's advocate on behalf of internalism.)

Each strategy builds on Lewis's (1979) observation that internalists must be relativists, in the sense that they must think that it is possible for a belief's content to have different truth values for different *subjects* and at different *times*. But each strategy goes much further than Lewis, representing a radical position hitherto unexplored in the literature on internalism.

The first strategy is to somehow use Lewis-style relativism to motivate *denying* TRANSPARENCY. (I'll leave it for those who wish to pursue this strategy to construct an argument from that form of relativism to the negation of TRANSPARENCY.) This strategy takes us into territory that is as yet unexplored by internalists, and it is even contrary to the spirit of mainstream relativism. As far as I know, no internalist so far has questioned TRANSPARENCY, and a good deal of work in relativist semantics is motivated by a desire to *keep* transparency-of-truth principles like TRANSPARENCY.[11] In the words of the leading contemporary relativist:

---

(2) $\vdash \Box((A \wedge B) \rightarrow C) \rightarrow (\Box(A \wedge B) \rightarrow \Box C)$

(3) $\vdash \Box(A \rightarrow (B \rightarrow (A \wedge B))) \rightarrow (\Box A \rightarrow \Box(B \rightarrow (A \wedge B)))$

(4) $\vdash \Box(B \rightarrow (A \wedge B)) \rightarrow (\Box B \rightarrow \Box(A \wedge B))$

The rest of the proof only requires truth-functional logic and a single application of (iii). Because $A \rightarrow (B \rightarrow (A \wedge B))$ is a closed FOL theorem, by (iii),

(5) $\vdash \Box(A \rightarrow (B \rightarrow (A \wedge B)))$,

(3) and (5) imply:

(6) $\vdash \Box A \rightarrow \Box(B \rightarrow (A \wedge B))$

(4) and (6) imply:

(7) $\vdash \Box A \rightarrow (\Box B \rightarrow \Box(A \wedge B))$

(7) implies:

(8) $\vdash (\Box A \wedge \Box B) \rightarrow \Box(A \wedge B)$

(1) and (2) imply:

(9) $\vdash \Box(A \wedge B) \rightarrow \Box C$

(8) and (9) imply:

(10) $\vdash (\Box A \wedge \Box B) \rightarrow \Box C$

By (10), $\{\Box A, \Box B\} \vdash \Box C$, so $\{\Box A, \Box B, \neg\Box C\} = \{\text{TRANSPARENCY}, \text{BROAD}_\text{T}, \text{NARROW}_\text{C}\}$ is inconsistent in the system.

[11] TRANSPARENCY is not a standard form or transparency principle, but, like the more standard principles, it underwrites 'disquotational' inferences—in this case, inferences from '*x* is a true belief whose content is *p*' to '*p*' (via the transparency of truth for contents).

> [E]ven committed relativists about some area of discourse will want the conveniences afforded by a disquotational [i.e., transparent] truth predicate when they are engaging in that discourse. Moreover, it is easy to give a semantics for monadic "true" and "false" that works in an analytic relativist framework and ratifies the disquotational inferences'. (MacFarlane 2011: 442)

Since MacFarlane is concerned with the semantics of language rather than thought, TRANSPARENCY is not among the transparency principles he discusses, but there is no greater difficulty involved in defining a transparent ('disquotational') truth predicate that applies to beliefs within a relativist semantic framework.[12]

The second strategy is even more radical than the first: it is to attempt to use Lewis-style relativism to motivate the view that TRANSPARENCY somehow does not make sense or is meaningless. An internalist pursuing this strategy will insist that the notion of relative truth for contents—truth relative to a subject and a time—is the only intelligible one, and it is simply nonsense to speak of a content being true or false *simpliciter*. This kind of internalist faces an uphill battle in explaining why MacFarlane has not, despite appearances, succeeded in deploying a meaningful monadic truth predicate.

Both strategies look seriously undermotivated, but that is not the main problem they have. The main problem is that they either lead to absurdities or deprive the internalist of the means to deny absurdities.

Consider the first strategy. Here is a definition of truth for beliefs: a belief is true iff its content is true. Here is a definition of truth for contents: the content that p is true iff p. The first strategy is committed to denying at least one of these definitions. Since they are definitions, we may necessitate them, and their necessitations entail TRANSPARENCY in the same weak logic assumed in §3, supplemented by propositional quantification, very elementary arithmetic, and bivalence.[13]

Obviously, the first strategy is also committed to the existence of counterexamples to TRANSPARENCY. But what would these counterexamples look like? They must be cases in which either (i) a subject S believes that p, and p, yet S's belief that p is false, or (ii) S believes that p, and it is not the case that p, yet S's belief

---

[12] I prefer 'transparent' to 'disquotational' because various transparency-of-truth principles do not involve quotation. TRANSPARENCY is a case in point.

[13] The necessitations of the definitions of truth for beliefs and contents are formalized, respectively, as (1) and (2).

(1) $\Box \forall x\ (x \text{ is true} \leftrightarrow \exists p(c(x) = p \wedge p))$

(2) $\Box \forall p\ (p \text{ is true} \leftrightarrow p)$

(3) is derivable from (1) and (2).

(3) $\Box \forall x\ (x \text{ is true} \leftrightarrow c(x) \text{ is true})$

Now given that 'x is true' is defined as $v(x) = 1$, and that we have necessitated bivalence ($\Box \forall x\ (v(x) = 1 \vee v(x) = 0)$) and $\Box 0 \neq 1$ as theorems, we can also derive TRANSPARENCY. Specifically, the system I have in mind is constructed exactly like the weak system described in §3, except in that the non-modal language we begin with has both first-order and propositional quantification, a propositional identity predicate (with the analogues of the first-order axioms for these), as well as $\forall x(v(x) = 1 \vee v(x) = 0))$ and $0 \neq 1$ as axioms.

(A technical aside: the above presentation cuts some corners. Since c must now be of type $\mathbf{e \rightarrow t}$, $v(x)$ and $v(c(x))$ cannot both be well-formed. In fact, we'll have to express truth value with two function symbols, v, v', of different types, and consequently we'll need two axioms of bivalence, and TRANSPARENCY will have the form $\Box \forall x\ v(x) = v'(c(x))$.)

that *p* is true. But this is absurd, as can be seen immediately by instantiating the variables. It is clearly not possible, say, that Jones believes that Antwerp is in Belgium, and Antwerp is in Belgium, yet Jones' belief that Antwerp is in Belgium is false. It is also clearly not possible that Jones believes that Antwerp is in the Netherlands, and Antwerp is not in the Netherlands, yet Jones' belief that Antwerp is in the Netherlands is true. The advocate of the first strategy is committed to the absurd position that cases like this are possible.

The advocate of the second strategy may seem to have it easier, since she will not be committed to the absurdities discussed above. But she is also not committed to the negations of those absurdities, since both the absurdities and their negations are inexpressible—are nonsense—according to her position. Claiming that an obvious truth is nonsense may be less of an affront to reason than denying one, but both of the relativist approaches are deeply uninviting.[14]

Bielefeld University
33501 Bielefeld, Germany
ylivakkuri@gmail.com

References

Farkas, K. 2006. Semantic internalism and externalism. In *The Oxford Handbook of Philosophy of Language*, ed. E. Lepore and B. C. Smith, 323-40. Oxford: Oxford University Press.
Farkas, K. 2008. *The Subject's Point of View*. Oxford: Oxford University Press.
Gettier, E. 1963. Is knowledge justified true belief? *Analysis* 23: 121-23.
Kaplan, D. 1989. Demonstratives. In *Themes from Kaplan*, ed. J. Almog et al., 481-563. Oxford: Oxford University Press.
Lewis, D. 1979. Belief *de dicto* and *de se*. *Philosophical Review* 88: 513-43.
MacFarlane, J. 2011. Simplicity made difficult. *Philosophical Studies* 156: 441-48.
Malmgren, A.-S. 2011. Rationalism and the content of intuitive judgements. *Mind* 120: 263-327.
Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131-93.
Searle, J. 1983. *Intentionality*. Cambridge: Cambridge University Press.
Segal, G. 2000. *A Slim Book About Narrow Content*. Cambridge, Mass.: MIT Press.
Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.
Williamson, T. 2013. Gettier cases in epistemic logic. *Inquiry* 56: 1-14.
Williamson, T. 2015. A note on Gettier cases in epistemic logic. *Philosophical Studies* 172: 129-140.
Williamson, T. 2017. Modal science. In Yli-Vakkuri and McCullagh 2017, 453-92.
Yli-Vakkuri, J. and M. McCullagh, eds. 2017. *Williamson on Modality*. London: Routledge.
Yli-Vakkuri, J. and J. Hawthorne. Forthcoming. *Narrow Content*. Oxford: Oxford University Press.

[14] I would like to thank Katalin Farkas, Kit Fine, Peter Fritz, John Hawthorne, Panu Raatikainen, Margot Strohminger, Lee Walters, Timothy Williamson, Jack Woods, an anonymous referee and an Associate Editor for *Analysis*, and audiences at the University of Leeds, the University of Oslo, the University of Tampere, the University of Salzburg, Umeå University, the University of Tartu, and the University of Bristol for helpful discussions and comments. This work was supported by the Alexander von Humboldt Foundation.