IF YOU BELIEVE YOU BELIEVE, YOU BELIEVE. A CONSTITUTIVE ACCOUNT OF KNOWLEDGE OF ONE'S OWN BELIEFS

Peter BAUMANN

ABSTRACT: Can I be wrong about my own beliefs? More precisely: Can I falsely believe that I believe that *p*? I argue that the answer is negative. This runs against what many philosophers and psychologists have traditionally thought and still think. I use a rather new kind of argument, – one that is based on considerations about Moore's paradox. It shows that if one believes that one believes that *p* then one believes that *p* – even though one can believe that *p* without believing that one believes that *p*.

KEYWORDS: self-knowledge, Moore's paradox, second-order beliefs

Can I be wrong about my own beliefs? More precisely: Can I falsely believe that I believe that *p*? Can I have a false second-order belief that I believe that *p* (where the belief that *p* is a first-order belief)? The question is whether a sentence of the following form can be true:

(1) S believes that he believes that *p*, but he does not believe that *p*.[1]

If all instantiations of the scheme (1) are false, then the following holds:

(2) If S believes that he believes that *p*, then he does believe that *p*.

In other words, all our second-order beliefs are true: $BBp \rightarrow Bp$.[2] This is the claim I will argue for.

However, *prima facie* it seems that it is possible to have a false second-order belief with the following content:

[1] For the sake of simplicity, I am not adding temporal indices except where clarity demands it. I assume here that S is attributing a belief to herself as a present one, not a past or future one.

[2] "B*p*" stands for "S believes that *p*." The scope of "B" is the narrowest possible one: B(B*p*) and B(*p*). I will omit parentheses in the following. The claim that $BBp \rightarrow Bp$ is (like some other claims here) one of necessity but I won't mention this below, just for the sake of simplicity.

© LOGOS & EPISTEME, VIII, 4 (2017): 389-416

(3) I believe that *p*.[3]

Why should the fact that someone believes something of the form of (3) entail anything about the truth of that belief? This idea runs against what many philosophers and psychologists have traditionally thought and still think.[4]

I will use a rather new kind of argument for the main thesis here, – one that involves considerations about Moore's paradox and amounts to a constitutive view of self-knowledge of one's beliefs.[5] The main argument will be developed in

---

[3] Here, (3) is meant as a report of a belief state, not as its expression (cf. Ludwig Wittgenstein, *Philosophical Investigations* (2.ed.) (Oxford: Blackwell, 1958), pp.190-192).

[4] See, e.g., among the psychologists: Richard Nisbett and Timothy D. Wilson, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (1977): 231-259; Richard Nisbett and Lee Ross, *Human Inference: Strategies and Shortcomings of Social Judgment*, (Englewood Cliffs: Prentice Hall, 1980), 195ff.; Timothy D. Wilson, "Strangers to Ourselves: The Origins and Accuracy of Beliefs about One's Own Metnal States," in *Attribution. Basic Issues and Applications*, eds. John H. Harvey and Gifford Weary (Orlando: Academic Press, 1985), 9-36; Alison Gopnik, "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences* 16 (1993): 9ff.; Daryl J. Bem, *Beliefs, Attitudes, and Human Affairs* (Belmont, CA: Brooks/ Cole, 1970); for the "anti-Cartesian" attitude against another transparency thesis see amongst philosophers Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2000), ch.4.

[5] In a very general way, I am inspired by a paper by Sydney Shoemaker (see his "Moore's Paradox and Self Knowledge," *Philosophical Studies* 77 (1995): 211-228, his "Moore's Paradox and Self-Knowledge," in Sydney Shoemaker, *The First Person Perspective and Other Essays* (Cambridge: Cambridge University Press, 1996), and also his "On Knowing One's Own Mind," *Philosophical Perspectives* 2 (1988): 183-209; see also David M. Rosenthal, "Self-Knowledge and Moore's Paradox," *Philosophical Studies* 77 (1995): 195-209, and Rogers Albritton, "Comments on Moore's Paradox and Self-Knowledge," *Philosophical Studies* 77 (1995): 229-239). He mainly argues that if S believes that *p*, then S believes or even knows he believes that *p*. I, however, argue for the converse claim (see also Byeong D. Lee, "Moore's Paradox and Self-Ascribed Belief," *Erkenntnis* 55 (2001): 359-370). Furthermore, Shoemaker's thesis is restricted to the case of rational people. Shoemaker's "Moore's Paradox" (1995): 225-226 also makes the converse claim, but much more tentatively, and with restriction to rational people (see with even more reservations, his "Moore's Paradox" (1996), 92, 93); the argument presented here does not rely on ideas about rationality. For a similar approach see Tyler Burge, "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society* 96 (1996): 91-116. The thesis that BB$p \rightarrow$ Bp is much stronger than Burge's earlier claim that "Cartesian" thoughts of the form "I am thinking the thought that water is wet" are always true (see his "Individualism and Self-Knowledge," *The Journal of Philosophy* 85 (1988): 649-663). Jaakko Hintikka, *Knowledge and Belief. An Introduction to the Logic of the Two Notions* (Ithaca, NY: Cornell University Press, 1962), 123ff. goes into a similar direction as I do here. He, however, does not rely upon Moore's paradox (even though he has a lot to say about it; see Hintikka, *Knowledge and Belief*, 64ff.). For more recent constitutive views which differ considerably from the proposal here see, e.g., Richard Moran,

sections 2-4. Section 5 discusses objections. But first I need to say more about the notion of belief and related notions in order to clarify the main thesis and set the stage.

1. Beliefs

I take beliefs to be dispositional mental states that can be both manifest and latent, – dispositions for occurrent thought and more indirectly also for behavior based on such occurrent thoughts.[6] Beliefs do not always express themselves in occurrent thoughts. In my dreamless sleep I still believe that 2+2=4 even though I am sleeping and not thinking at all about numbers. One of the characteristics that distinguish beliefs from other mental states is a specific relation to truth: Their contents are held true by the subject. Desires and other mental states are different in that respect. Beliefs are "cognitive" in this sense; one could also say that a belief is a cognitive attitude towards some content.

In the case of a self-attributing second-order belief the notion of "I" lies within the scope of the second-order belief; it is not sufficient for such beliefs to attribute a belief to someone who happens to be me if I don't think of that person as myself. We are dealing with *de dicto*-beliefs about oneself here,[7] not with *de re*-beliefs. Similarly, the notion of a belief, too, lies within the scope of the second-order belief. If somebody ascribes a belief to herself, then she must be clear about the type of attitude she ascribes to herself. One cannot, for conceptual reasons, believe that (3) is true of oneself and not believe it is a *belief* (that *p*) that one has here. Believing the latter presupposes that one possesses the concept of a belief and that one knows certain basic things about beliefs. One need not have a psychological theory of belief but one needs to know, say, that there is a difference between beliefs and other kinds of attitudes (like desires, for instance). If one does not know these basic things then one does not possess the concept of a belief and thus cannot have second-order (*de dicto*) beliefs.[8] All this will become important below.

Authority and Estrangement. An Essay on Self-Knowledge (Princeton: Princeton University Press, 2001), Fordi Fernández, "Self-Knowledge, Rationality and Moore's Paradox," *Philosophy and Phenomenological Research* 71 (2005): 533-556, and Mathiey Doucet, "Can We Be Self-Deceived about What We Believe? Self-Knowledge, Self-Deception, and Rational Agency," *European Journal of Philosophy* 20 (2012): E1-25.

[6] If not indicated otherwise, I will use "thought" for "occurrent thought."

[7] One could add: with *de se*-beliefs (see David Lewis, "Attitudes de Dicto and de Se," *Philosophical Review* 88 (1979): 513-543).

[8] This holds even given an externalist account of mental or semantic content.

To avoid misunderstandings: By "second-order beliefs" I do not mean beliefs that one does not have a certain first-order belief. It is certainly possible for people to have repressed beliefs that they think they don't have. Jack might just laugh at the thought that his parents abandoned him when he was 4 years old but psychotherapy might uncover that he has a repressed belief that this was indeed the case. This example is of the following form:

> S believes that he does not believe that *p*, but he does believe that *p*.

This is certainly possible but I am not dealing with this case here (see also section 5.1 below).[9]

2. The Argument: First Part

Suppose that

> (4) S believes at t-1 that he believes that *p*.[10]

What does this entail?
Dispositional beliefs are often latent. However, there is a condition on dispositional beliefs which seems very plausible:

> (5) If S believes at t-1 that *p*, then S manifests that belief as an occurrent belief at t* (which is either before or at t-1).[11]

A few remarks on (5) are necessary before we can make the next step in the argument. - The idea behind (5) is that one cannot believe, say, that dogs bark without ever manifesting that belief up to then, that is, without ever occurrently thinking and holding true (up to then) that dogs bark. It doesn't matter whether the thought is a conscious or an unconscious, or even a "Freudian" one (where a thought is "unconscious" in case the person is not aware of having the thought, and "Freudian" if the person cannot (easily) become aware of it). I don't see any reason to deny that one can have unconscious occurrent thoughts; otherwise each thought

[9] See Shoemaker, "Moore's Paradox" (1995), 226; cf. against this Byeong D. Lee, "Shoemaker on Second-Order Belief and Self-Deception," *Dialogue* 41 (2002): 279-289.

[10] In (4) as well as in the antecedent of (5) "believes" is used in the full dispositional sense, covering both latent and manifest belief.

[11] A related reverse principle might seem more uncontroversial: If S has an occurrent belief at t that *p* then S has a dispositional belief at t that *p*. The dispositional belief might be as short-lived as an occurrent belief which comes and goes. But the disposition is still there as long as the occurrent thought is there. Something made the subject think that *p*, and if the same conditions were to hold again the subject would think again that *p* – even if, as a matter of fact, these conditions never come up again.

would be accompanied by the thought that one is having it.

To be sure: One can have a disposition to form a certain belief and for this disposition one does not need any antecedent manifest thought with the content of that belief. Jack might have never thought about the question whether zebras in the wild wear raincoats.[12] He might not have a belief that they don't (nor any alternative view on the matter). However, we can still assume that he has a disposition to form the belief that zebras don't wear raincoats in the wild; for instance, when asked whether they do he might well form such a belief (see Robert Audi's useful distinction between dispositional beliefs and dispositions to believe;[13] Audi, however, holds that one can have a dispositional belief without ever having had the corresponding occurrent belief). Another example: Someone can have a dispositional belief that he is 6 feet tall. Even given knowledge that 6 feet is much less than 12 miles, it does not follow that the subject has a dispositional belief that he is less than 12 miles tall. However, we might very well need to ascribe a disposition to form the relevant belief to the subject. If we gave up on the distinction between dispositions to believe and dispositional beliefs we would get an "inflation of beliefs" and would have to attribute implausibly many beliefs to subjects. For instance, there are many propositions which a person does not accept in the present but will come to accept in the future; this is a basic fact of life. If, as seems very plausible, the future acceptance of some proposition results from the triggering of a relevant prior disposition, and if that disposition is not a disposition to believe but rather a dispositional belief then the person would already count as believing a proposition way before they accept it. This seems very odd and strongly suggests that we need to distinguish between dispositions to believe and dispositional beliefs.[14] Here is another way to mark the difference. Dispositional beliefs are first-order dispositions of thought and subsequent behavior while dispositions to believe are second-order dispositions to develop and

[12] See Daniel Dennett, "A Cure for the Common Code?" in his *Brainstorms. Philosophical Essays on Mind and Psychology* (Cambridge, MA: MIT Press, 1978), 90-108, especially 104; I am using Dennett's example contrary to his own purposes.

[13] See Robert Audi, "Dispositional Beliefs and Dispositions to Believe," *Noûs* 28 (1994): 419-434, and especially 420-421.

[14] Here is one more example. Does a newborn baby believe that there is no greatest prime? If there is no difference between dispositional beliefs and dispositions to believe then it won't be easy to deny this kind of belief to newborns. Sure, the baby doesn't have the notion of a number yet but given the right circumstances (including normal development) it will acquire it plus the belief that there is no greatest prime. So, we need to draw a line somewhere between dispositional beliefs and dispositions to believe – whether we use these expressions or others. The claim (5) above formulates a very plausible criterion to that effect.

form such first-order dispositional beliefs. This distinction is very important und useful in the case of cognitive attitudes. The first-order disposition, the dispositional belief, is a cognitive attitude towards some content while the second-order disposition to form such a belief does not involve any cognitive attitude towards that content (though, perhaps, cognitive attitudes towards other contents).[15]

But isn't it possible that a belief manifests itself in action but not in thought? Can't there be "unthought beliefs" driving our behavior? I don't think so. First, as explained above, I take beliefs to be cognitive attitudes and states. A merely behavioral disposition to act or behave as if *p* without there being or having been any kind of occurrent thought that *p* in the subject's mind is not a cognitive state. It is therefore not clear at all why one should call such a state a "belief" (one can certainly redefine terms as one likes but in this case this would be misleading). Second, as pointed out above, an occurrent belief need not be conscious: One can have it without being aware of having it. Hence, that there has been no conscious occurrent belief at t* does not mean that there hasn't been any occurrent belief at t*. One should not mistake the presence of an unconscious (occurrent) belief for the lack of (occurrent) belief altogether.[16] Third, as pointed out above, one needs to take the distinction between dispositions to believe and dispositional beliefs very seriously; a mere disposition to form some belief that *p* does not constitute a belief that *p*. Finally, even if one still has doubts about (5) one should keep in mind that I am only dealing with the application of (5) to the case of second-order beliefs here (see below). Even if one could make sense of "unthought belief" in general, it would still be very hard to imagine how it should be possible that a second-order belief can express itself in action but not in thought. Can Jack believe that he has the belief that he is good looking but never manifest that second-order belief in thought but only in action? In what kind of action? I doubt there are any "real life"-explanations of behavior in terms of unthought higher-order beliefs. To assume that there can be unthought second-order beliefs (in contrast to unthought first-order beliefs) thus seems very hard to justify.

Now, (4) and (5) entail

---

[15] Interestingly, there does not seem to be that much of a place for this distinction in the case of non-cognitive dispositions. The behavior of sugar cubes in water can be very usefully described and, perhaps, even explained in terms of the first-order disposition of water-solubility of sugar; there does not seem to be much need for notions of second-order dispositions here; it seems talk about second-order dispositions can be easily replaced here by talk about first-order dispositions.

[16] That all occurrent beliefs are conscious is a very strong claim anyway. One might wonder whether there is a threat of an infinite regress here (is the subject conscious that they are conscious that…?). Apart from that: What counts in favor of such strong "Cartesianism"?

(6) S occurrently believes at t* that he believes that *p*.

A necessary condition for belief becomes important now: a condition sometimes called a condition of "minimal rationality."[17] Here is an example.[18] Someone who thinks that McKinley has been assassinated cannot ignore (that is, not spend any thought on or be unaware of) the question whether McKinley is dead. This does not mean that the subject needs to have the notion of a question or must ask himself the question explicitly whether McKinley is dead ("Hey, is McKinley dead?"). Rather, the subject must somehow be aware of the issue whether McKinley is dead. In the case of this example, the subject must also grasp and agree with the idea that McKinley is dead. Otherwise, he does not even count as someone who believes that McKinley has been assassinated. For analogous reasons, I want to argue in a moment, someone who thinks and believes that he believes that *p*, cannot ignore and must be aware of the question whether *p*. He must have some thought (though not necessarily a positive view one way or another) about that question, too. Otherwise, he wouldn't even count as someone who believes that he believes that *p*. He just wouldn't be grasping the concept of a belief. But the latter is, as we have seen above, necessary for having second-order beliefs.

Thinking that I believe that *p* is a way of thinking about *p;* here one thinks about *p* as something towards which one can have certain attitudes like belief or disbelief. It is thus *ipso facto* a way of being aware of the question whether *p*, given that such attitudes can only be thought of as ways of settling questions.[19] To put it differently: Thinking that I believe that *p* is about the particular propositional attitude of belief. One can think of belief only as something that settles a question. Thus one is then *ipso facto* aware of the question. In other words: One doesn't ignore the question. Therefore, one cannot think that one believes that *p* without thinking in some form that there is a question as to whether *p*; without the latter one's own thought would not be intelligible to oneself (per impossibile). Sure, one does not have to have a conscious belief of a form like "There is the question as to

[17] See, e.g., Christopher Cherniak, *Minimal Rationality* (Cambridge, MA: MIT Press, 1986). To call it that can be a bit misleading. A subject who fails to meet that kind of condition fails to have or fails to be able to have certain concepts and beliefs. However this does not mean that such a subject is irrational (not even minimally) but simply that it does not have what one needs in order to master a given concept and entertain certain beliefs involving those concepts.

[18] See Stephen Stich, *From Folk Psychology to Cognitive Science: The Case against Belief* (Cambridge, MA: MIT Press, 1983), 56.

[19] Settling a question can also be very easy and even trivial, like, e.g., in the case of 1 plus 1 equaling 2.

whether *p* and I have answered it by endorsing *p* so that I believe that *p*" when one thinks that one believes that *p*. Similarly, one does not have to have a conscious belief of a form like "McKinley is dead" when one thinks that McKinley has been assassinated. But as someone who thinks that McKinley has been assassinated cannot and does not ignore (is aware of) the question whether he is dead, so someone who thinks that he believes that *p* cannot and does not ignore (is aware of) the question as to whether *p*. It is important to stress that awareness comes in degrees; a subject need not be maximally aware of a question in order to be aware of it. Something does not need to be in the central focus of one's awareness; it could be closer to the periphery of awareness but the subject would then still count as being aware of it.

Hence, given (6) and given that having beliefs requires this kind of awareness we have to accept

(7) S does not at t* ignore the question whether *p*.

Since S has a thought (even if it is only the thought or awareness that there is a question here; thoughts need not take the form of explicit deliberation) about the question whether *p* insofar as he thinks that he believes that *p*, we can put (6) and (7) together and say that

(8) S occurrently believes at t* that he believes that *p*; this involves thoughts about the question whether *p*.

If one does not ignore or if one is aware of a question, then one has thoughts about it.

But does S, in addition, have to believe that *p*? What kinds of attitudes can S have towards *p* when he occurrently believes that he believes that *p* and is aware of the question whether *p*? More precisely: What kinds of attitudes can S have towards *p* when he is aware of and has thoughts about the question whether *p*? There are exactly three options, three stances the subject can take or have (with no other alternative option):

i. the positive stance: to believe occurrently that *p*,

ii. the negative stance: to believe occurrently that not *p*, or

iii. the indifference stance: to leave it open (occurrently) whether *p*.

Leaving it open is a residual category here. It involves everything between simple lack of a positive view one way or the other about (but still involving awareness of the question) whether *p* on the one hand and explicit suspension of

belief about whether *p* on the other hand.[20]

Hence, S – who occurrently believes at t* that he believes that *p* – can only be in one of the following three situations (at that time):

> I. occurrently believe at t* that he believes that *p* and occurrently believe at t* that *p*,
>
> II. occurrently believe at t* that he believes that *p* and occurrently believe at t* that not *p*, or
>
> III. occurrently believe at t* that he believes that *p* and (occurrently) leave it open at t* whether *p* (while being aware of the question whether *p*).

It appears then that either S has pairs of thoughts here - a second-order and a first-order thought – or S has a pair of a second-order thought and an indifference stance on the content of the relevant first-order belief. Let us look at pairs of thoughts, first (I, II). Since S has the thought about the question whether *p* in the context of the second-order thought that he believes that *p*, it is plausible also to ascribe a single complex conjunctive thought to S. The following (schema of a) conjunction principle is plausible:[21]

> (Conj-1) If S believes occurrently that he believes that *p*, is aware of the question whether *p*, and takes a positive or negative stance on *p*, then S has an occurrent belief in the conjunction of the contents of the second-order occurrent belief and of his positive or negative stance on *p*.[22]

S thus is in one of the following three situations (brackets indicating the content of the thought): He

---

[20] Is the indifference stance a second-order attitude or does it involve a second-order belief that one does not have a first-order belief about *p*? One may call such a second-order belief an "indifference stance" but what I have in mind here need not be and usually isn't of the second order, like the state of being epistemically indecisive about whether *p*. Young children or non-human animals might not be able to have higher order attitudes because they lack concepts like *belief* but they can be indecisive about something. Also, one can be indecisive concerning options for choice even if one does not assume a higher-order attitude; how then could there not be a parallel in the case of belief?

[21] See, e.g., Simon Evnine, *Epistemic Dimensions of Personhood* (Oxford: Oxford University Press, 2008), chs. 3-4 for a defense of such principles.

[22] One can argue that (Conj-1) as well as similar conjunction principles hold for dispositional, latent as well as manifest, beliefs, including also unconscious or even "Freudian" beliefs; however, I cannot go into the details of the different cases here. – Other conjunction principles are simpler and more straightforward but false, like the following one: If S believes that *p* and also believes that *q*, then S believes that (*p and q*). More plausible is a principle of conjunction elimination: If S believes that (*p* and *q*), then S believes that *p* and S believes that *q*.

> 1. occurrently believes at t* that (*p* and he believes that *p*),
>
> 2. occurrently believes at t* that (not *p* and he believes that *p*), or
>
> III. occurrently believes at t* that he believes that *p* and leaves it open whether *p* (while being aware of the question whether *p*).

Now, what about III? It makes the antecedent of (Conj-1) false and the principle irrelevant here. But is there, perhaps, another conjunction principle for pairs of beliefs and indifference stances? An indifference stance towards some proposition *p* does, of course, not express itself in the belief that *p* or the belief that not *p*. Hence, there is no direct parallel to 1. and 2. above. It would be nonsense to say something like the following:

> A subject in condition III occurrently believes at t* that (*???* and he believes that *p*).

However, there is a less direct parallel to 1. and 2. above which is direct enough for our purposes here:

> (Conj-2) If S believes occurrently that he believes that *p*, is aware of the question whether *p*, and takes an indifference stance on *p*, then S has a conjunctive occurrent belief of the form "I believe that *p* but the question whether *p* is unsettled for me".[23]

Hence, we also get from III and (Conj-2) to the claim that S

> 3. occurrently believes at t* that (it is unsettled whether *p* and he believes that *p*).

So, from I, II, III plus both conjunction principles we get to the conclusion that our subject can only be in condition 1., 2. or 3. Now, for conceptual reasons S can only be in situation 1. Why? Let us take situation 2 first. Someone who were in that situation would occurrently believe at t* something of the form "I believe that *p*, but not *p*." This is a Moore-paradoxical thought.[24] The problem with that[25] is not

---

[23] Again: Awareness admits of degrees and something can be more or less in the focus of one's awareness.

[24] See George Edward Moore, "Russell's Theory of Descriptions," in his *Philosophical Papers* (London/ New York: Allen & Unwin/ Macmillan, 1959), 151-195, especially 175-176.

[25] I am only using commissive versions of Moore's paradox here (of the form "I believe that *p* but not *p*"); ommissive versions ("*P* but I don't believe it") are irrelevant to my argument. See also section 5.1 below. – I am leaving aside uses of such phrases by eliminativists about belief: "*p* but since there is no such thing as belief I don't believe it!" (see, e.g., Paul M. Churchland, "Eliminative Materialism and the Propositional Attitudes," *The Journal of Philosophy* 78 (1981): 67-90). Eliminativism about belief is not incoherent or Moore-paradoxical. A Moore-paradoxical thought or utterance presupposes that there are beliefs while eliminativists about belief deny that and are not even in a position to make a truly Moore-paradoxical statement. – Even though

just that it would be an incoherent thought.[26] No, one cannot even have such a thought.[27] Why not? Here we can use an important remark by Wittgenstein:[28] One can mistrust one's senses but one cannot mistrust one's own belief.[29] Having a belief that *p* is incompatible with having a second-order attitude of mistrust towards that belief: for instance, holding that the belief that *p* is false (or suspending judgment on the question whether it is true: see below). However, a subject who is in situation 2 above and thinks and believes something of the form "I believe that *p*, but not *p*" would have such a second-order attitude of mistrust towards his belief. Since the latter is not possible, the former is also not possible (both for conceptual reasons). If someone said or thought something of the form "I believe that *p*, but not *p*," then whatever he was ascribing to himself it couldn't be a belief; hence, he couldn't thereby express or manifest a second-order (manifest) belief; his use of the word "belief" (his attempt at tokening of concept of belief) would show that he hasn't yet mastered the concept of belief. In other words, we can exclude situation 2 as impossible here.[30]

Moore's paradox is often discussed as a problem of assertion, it has long (even before Wittgenstein, *Investigations*) been recognized that the same problem arises for thought not expressed linguistically.

[26] – or an absurd thought; see Uriah Kriegel, "Moore's Paradox and the Structure of Conscious Belief," *Erkenntnis* 61 (2004): 99-121.

[27] Shoemaker, "Moore's Paradox" (1995), sec. IV and Shoemaker, "Moore's Paradox" (1996), sec.IV agree but for different reasons than those presented here.

[28] See Wittgenstein, *Investigations*, 190; more on that also below.

[29] But cf. also Béla Szabados, "Wittgenstein on Mistrusting One's Own Belief," *Canadian Journal of Philosophy* 11 (1981): 603-612.

[30] One could speculate about the possibility of a "division" of the mind and consider cases where one "sub-subject" disagrees with what another "sub-subject" believes. One sub-subject might hold that it believes that *p*, while the other sub-subject might hold that *not p* (compare this to Wittgenstein *Investigations*, 192: "If I listened to the words of my mouth, I might say that someone else was speaking out of my mouth"). However, such deviant cases can be left aside here: They are cases of "split minds" and not ordinary cases of self-attributing beliefs which are topical here. But one might object to this that it makes sense to say something like "Spiders are harmless but when I think about my behavior when I'm near a spider I come to the conclusion that I still believe that spiders are not harmless" or, shorter, "Spiders are harmless but I still believe they aren't"? This makes some sense but it is crucial, again, to acknowledge that such a subject identifies only with part of her mind and treats her behavior as if it were someone else's. Strictly speaking, such a belief is not the subject's belief but the belief (if one may use this word here) of some sub-personal agent or module (see for this also Stephen Stich, "Beliefs and Subdoxastic States," *Philosophy of Science* 45 (1978): 499-518); the fact that some sub-personal agent or module holds a belief (or an attitude like that) does not entail that the person herself holds that belief (compare this with the bad inference from the claim that a particular group

To be sure, this does not mean or imply that one couldn't be less than perfectly confident in one's beliefs (have an intermediate degree of belief), imagine the possibility of being wrong,[31] see one's evidence for one's belief as imperfect, etc. However, all this does not amount to mistrusting one's beliefs.

All this entails that the verb "falsely believe that *p*" has no use in the first-person, present tense.[32] As we will see, only the verb "truly believe that *p*" does. This might explain why we usually skip the qualification "truly" when self-attributing beliefs. False beliefs that oneself might have are "blindspots"[33] in the sense that they are not self-attributable as false beliefs. Having a false belief is an essentially "intransparent" condition insofar as the person cannot know or even believe that he is in this condition while he is in it.

For reasons analogous to the ones above, the situation 3 also turns out to be impossible. Someone who were in that condition would occurrently believe at t* something of the form: "I believe that *p* but it is unsettled whether *p*." This also expresses an attitude of mistrust towards one's own belief that *p* (though a softer one). But one cannot take such an attitude of mistrust towards one's own beliefs. Hence, we can – for similar reasons to the ones concerning situation 2 – exclude situation 3 as impossible (also for conceptual reasons).

But if cases 2 and 3 are excluded as impossible, then only case 1 remains – and there is nothing problematic or incoherent about that one. Hence, we can conclude from the above remarks about cases 1-3 and (8) that

> (9) If S occurrently believes at t* that he believes that *p*, then S occurrently believes at t* that he believes that *p*, and *p* (in the sense of *"p* and I believe that *p*").

There is a plausible principle of distribution of belief over conjunction:[34]

> (Dist) If S believes that (*p and q*), then S believes that *p*.

Given (Dist) we can move from (9) to

> (10) If S occurrently believes at t* that he believes that *p*, then S occurrently believes at t* that *p*.

---

member holds a certain belief to the claim that the group holds that belief or view).

[31] Concessive self-attributions of beliefs ("I believe it's going to rain but I could, of course, be wrong about that") do not constitute cases of mistrust of one's belief: the confidence that one is right can still be quite firm.

[32] See Wittgenstein, *Investigations*, 190.

[33] See also, more generally, Roy A. Sorensen, *Blindspots* (Oxford: Clarendon, 1988).

[34] See, e.g., John N. Williams, "Wittgenstein, Moorean Absurdity and its Disappearance from Speech," *Synthese* 149 (2006): 225-254, especially sec.7.

3. More on Mistrusting One's Beliefs

Before we move on with the argument, some more remarks about why one cannot mistrust one's own beliefs seem useful. Let us look at the clearest case of a Moore paradoxical belief (similar arguments can be made for other forms of Moore-paradoxicality, like holding that one believes that *p* but suspending judgment about whether *p*): the (alleged) belief that

(MP) Not *p*, but I believe that *p*.

Why can one not believe something of the form (MP)?

I can mistrust another person's belief. In such a case I hold, so to speak, my belief against another person's belief (or against what I take to be her belief). I compare them and if there is disagreement (given that I can see no reason to revise my own belief), I go with my own belief. Why with mine? Well, that is what it means to have a belief: One goes with it (against alternative beliefs, given that the fact of disagreement or related facts do not itself give one a reason to change one's belief). If I check other people's beliefs, I cannot but use my beliefs as the standard (even if I originally got my beliefs from others and even if I change my beliefs under the influence of other people's beliefs). Sometimes – like in the case we're focusing on here – I have an explicit belief about the relevant subject matter ("He thinks it's raining but it isn't"). But at other times I don't: there might just be a reluctance to judge the whole thing ("He thinks it's raining but that's not clear at all"). This reluctance, however, is also based on certain beliefs ("He doesn't know the weather conditions," "Conditions of perception are much too bad to judge this," etc.). In both cases, I have a belief or a set of beliefs that is the basis for mistrust towards another person's belief.[35]

I cannot do anything like that in my own case. I would have to treat myself as if I were not myself but another person. Given that in (MP) I would have to think of myself as myself ("I"), I would have to think of myself as myself and as another person. However, mastering the notion of oneself as well as the notion of others involves knowing that oneself is not another person different from oneself. Whoever says something to the effect of "I am not myself" is either not sincere or uses language in a special way or just documents that he has not mastered words like "I" and "someone else." Mastering such notions is a precondition of being able

[35] Couldn't a non-propositional mental state be the basis for my mistrust of someone's belief? Suppose sensations, for instance, are such non-propositional states. But in what sense could they be a basis for my mistrust if they don't lead to certain beliefs which then form the more immediate basis for my mistrust? This touches on a whole series of questions which cannot be pursued in further detail here.

to have second-order beliefs. Hence, one cannot have thoughts or beliefs of the form (MP). I cannot hold my own beliefs against my own beliefs as an (allegedly) independent standard;[36] I cannot mistrust my own beliefs. In other words, when I say that S holds something false true I imply (or implicate) that something that is acknowledged by me is not acknowledged by S. I am thus assuming that there is an epistemic asymmetry between me and S which explains why S is wrong and I am right. This only makes sense given the assumption that I am not S. Since I cannot take myself to be S (not me), I cannot apply the above asymmetry to my own case. In other words, I cannot take myself to hold something true that I think is false.[37]

Hence, the subject cannot mistrust his own beliefs and believe something of the form of

(MP) Not *p*, but I believe that *p*.

Somebody who sincerely claims to believe such a thing only shows that he has not mastered the concept of belief. Even if one were to argue that he expresses some kind of second-order belief, it wouldn't and couldn't be one with the content (MP). Hence, (MP) cannot express a self-ascription of a belief. Not that it constitutes an irrational or defective self-ascriptions of a belief; rather, (MP) does not express any possible self-ascription of a belief at all.

4. The Argument: Second Part

Back to

(10) If S occurrently believes at t* that he believes that *p*, then S occurrently believes at t* that *p*.

(10) is not our thesis, even though quite close: It does not say that (omitting the reference to a given point in time)

(2) If S believes that he believes that *p*, then he does believe that *p*

– where "belief" is used in the wide sense including non-manifest, merely dispositional belief as well as occurrent belief. Can we generalize (10) to include situations in which S believes that he believes that *p* but only in a latent and non-manifest way? Can we generalize (10) such that it entails (2)?

First a brief reminder. Suppose that S, at t-2, believes (in a merely dispositional, that is, latent sense) that he believes that *p*. According to (5) as

[36] See Wittgenstein, *Investigations*, 190.

[37] Suppose I have changed my mind: Yesterday I believed that *p* but today I believe that not *p*. Then, I can say today I was wrong yesterday. However, this does, of course, not amount to saying that my present belief is false (see also the remarks above in section 1).

applied to this case and to some earlier time t-1, the following holds

> If S believes latently at t-2 that he believes that *p*, then S manifests that belief as an occurrent belief at some earlier time (which we can call "t-1" here).

Hence, given our assumptions about S here, S manifests the belief that he believes that *p* at t-1. We can now use the relation between dispositional, latent and manifest belief to show that (2) is true if (10) is.

The crucial point here is that if the occurrent belief of S at t-1 that he believes that *p* leads to the latent belief of S at t-2 that he believes that *p*, then the occurrent belief of S at t-1 that *p* – which comes with the corresponding occurrent second-order belief (see (10) above) – will also lead to the latent belief of S at t-2 that *p*. In other words, the same occurrent second-order belief at t-1 leads to both corresponding second- and the first-order latent beliefs at t-2. In a nutshell: The latent second-order belief can only arise from circumstances which also give rise to the corresponding latent first-order belief.[38] The first comes with the second (for a worry, see below).

Here is a different way to put it. A latent belief in some proposition is a disposition to, amongst other things, think that proposition (given certain triggering conditions). The latent belief that one believes that *p*, for instance, is a disposition to think that one believes that *p*. Since one cannot (see (10)) occurrently think that one believes that *p* without also occurrently thinking that *p*, the same disposition (given the relevant circumstances) triggers the thought that one believes that *p* as well as the thought that *p*. Hence, this very disposition is also a disposition to think that *p*. In other words, if S has a latent belief that he believes that *p*, then S also has a latent belief that *p*:

> (11) If S latently believes at t* that he believes that *p*, then S cannot but latently believe at t* that *p*.

Since beliefs are either manifest or latent, we can put (10) and (11) together and thus get our core thesis (again, skipping temporal indices for the sake of simplicity):

> (2) If S believes that he believes that *p*, then he does believe that *p*.

But, one might ask incredulously, isn't it possible that S, after t-1, continues to believe that he believes that *p* (though in a latent way) but loses the belief that *p*? Just stopping to think about *p* would not be sufficient for that: S has at t-1

[38] If the latent first-order belief already exist independently and antecedently, then there is overdetermination and the second-order belief merely "reconfirms" the first-order belief (see also the third-to-last paragraph in this section). This does not constitute a problem here.

acquired (or reconfirmed) the dispositional belief that *p*. As long as S doesn't change his mind about *p*, we can still attribute the belief that *p* to him. But couldn't S change his mind about *p*? And at the same time stick with his latent belief that he believes that *p*?

Not according to the view defended here. Suppose S changes his mind at t-2 about *p*: He now, e.g., comes to believe that not *p*. The critical assumption (for reduction) is that he still has, at t-2, his dispositional belief that he believes that *p*. Now, as we just saw: If at t-2 S has this belief, then he also has the dispositional belief that *p*. So, our situation would be rather one where the subject has inconsistent beliefs: one belief that not *p*, and another belief that *p* (this differs, of course, from the case of holding a belief that *p* and not *p*). It is not clear whether one can describe this as a change of mind but certainly S has thus not lost his belief that *p*.

But couldn't S change his mind at t-2 in a different way: not by acquiring the belief that not *p* but by simply losing the belief that *p*? Again, our assumption is that he still has, at t-2, the dispositional belief that he believes that *p*. Hence, he would (see above, again) also have the dispositional belief that *p*. The assumption that the subject has just dropped a belief thus leads to an inconsistency not of the beliefs of the subject but in the description of the subject's situation: as both having and not having the belief that *p*.

Our subject thus cannot be in a different mind about *p* without changing his mind about whether he believes that *p*. (2) remains standing and we can conclude that

$$BBp \rightarrow Bp.$$[39]

The argument for (2) has interesting consequences. If I assent (mentally or linguistically) to *"p*, and I believe that *p*", then I cannot detach the second conjunct and leave the first part behind, so to speak. Others can separate the two "parts" when they think or talk about me: They might think that I believe that *p*, but they need not hold that *p*. From the first-person perspective everything is different:

[39] See, though not quite in agreement with the argument above: Christopher Peacocke, *A Study of Concepts* (Cambridge, MA: MIT Press, 1992), 158, Tom Stoneham, "On Believing that I Am Thinking," *Proceedings of the Aristotelian Society* 98 (1998): 125-144, and U.T. Place, "The Infallibility of Our Knowledge of Our Own Beliefs," *Analysis* 31 (1970/71): 197-204; cf. against that Hugh Mellor, "Conscious Belief," *Proceedings of the Aristotelian Society* 78 (1977/78): 88-101, especially 91f. – I do not rule out that one can believe that not all of one's beliefs are true; the Preface Paradox is not Moore-paradoxical (though related). – The above argument works for full belief; I think a similar argument (though much more complicated in detail) can be made for degrees of belief but I will not attempt this here.

From this perspective *"p*, and I believe that *p*" is not a normal conjunction insofar as I cannot infer "I believe that *p*" from it without committing to *p* at the same time.[40]

A second-order belief is "constitutive" of the corresponding first-order belief. "Constitutive" is meant in a conceptual sense here, not in an empirical sense. The second-order belief brings with it, "involves" the first-order belief, and all that for conceptual reasons.[41] If S at t acquires a second-order belief that he believes that *p* and if S did not, before t, believe that *p*, then she acquires the belief that *p* just because she acquires the second-order belief that she believes that *p*. She might, of course, already believe that *p* before her acquisition of the relevant second-order belief, then think about whether or not *p* and about her views on whether or not *p*, and thus finally come to acquire her second-order belief that she believes that *p*. In this case, the second-order belief does not create the first-order belief but rather "reconfirms" it (see fn.38). It is also possible that the acquisition of a second-order belief creates an inconsistent mind set. Suppose S believes that not *p*. Suppose also that she somehow acquires the belief that she believes that *p* (a clever psychiatrist might convince her that he does). Then she thereby also acquires the belief that *p* – which is inconsistent with her belief that not *p*.[42]

I have only argued for a conditional thesis here (2). Hence, insofar as (2) leaves it open whether we do indeed have second-order beliefs, it is also left open whether we have any true beliefs about our own beliefs. It is left open whether we have any self-knowledge about our own beliefs. Now, it might be the case that one cannot have beliefs without having at least *some* second-order beliefs; that we have first-order beliefs would entail that we also have second-order beliefs. However, I want to leave *that* open here.[43] I would rather assume that, as a matter of fact, we often do have second-order beliefs (whatever the explanation of this fact is). Given what I have just said, this would entail that we are indeed right

---

[40] See André Gallois, *The World without, the Mind within. An Essay on First-Person Authority* (Cambridge: Cambridge University Press, 1996), 5-7, 46, passim who argues that questions about *p* and questions about my beliefs about *p* are not separate when raised from the perspective of the first person.

[41] See Crispin Wright, "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy, and Intention," *The Journal of Philosophy* 86 (1989): 622-634; Jane Heal, "On First-Person Authority," *Proceedings of the Aristotelian Society* 102 (2002): 1-19.

[42] See Derek Bolton, "Self-Knowledge, Error and Disorder," in *Mental Simulation: Evaluations and Applications*, eds. Martin Davies and Tony Stone (Oxford: Blackwell, 1995), 209-234, and Shoemaker, "Moore's Paradox" (1996), 89-91. See also section 5.5. below.

[43] See, e.g., Donald Davidson, "Rational Animals," *Dialectica* 36 (1982): 317-327.

about our own beliefs when we think about it.[44]

No person needs to know or have true beliefs about all her present first-order beliefs, perhaps not even about any of them. Even if error, the presence of a false belief, about one's own beliefs is impossible, ignorance, the absence of a true belief, is still possible. If S has the belief that *p* he need not have the second-order belief that he believes that *p*. Not: B*p* → BB*p*.[45]

5. Objections

Finally, I would like to consider and reply to some objections.

5.1. Believing that One Doesn't Believe

Let's start with what is perhaps the most serious objection I can think of. As I have already mentioned above (section 1), I am only dealing with beliefs that one has a belief that *p* (BB*p*), not with beliefs that one does not have a belief that *p* (B not B*p*). I don't see any convincing argument similar to the one above that would support the following claim: B (not B*p*) → not B*p*. Such a claim would easily lead to a contradiction. Suppose the subject has a suppressed and unconscious belief that *p* (B*p*). Somebody (a clever psychoanalyst for example) convinces her that she does not believe that *p* (B not B*p*). If "B (not B*p*) → not B*p*" were true, a contradiction would follow: B*p* & not B*p*.

But isn't there an argument for "B (not B*p*) → not B*p*" which is parallel to the one above for "BB*p* → B*p*"? And if the former leads to a contradiction, how then can we still hold on to the latter? Here is the idea (see sections 2-4 above for the details of the parallel). If S believes at t-1 that he doesn't believe that *p*, then he believes occurrently at some earlier time t* that he doesn't believe that *p*, is aware of the question whether *p* and thus has some thought and stance about whether *p*. Given that S can only have the three stances towards *p* mentioned above, S must

[44] How can I move from the claim that our second-order beliefs are true to the claim that they constitute knowledge? Couldn't some true second-order beliefs fail to be knowledge? I don't see how this should be possible – given the type of argument above. However, I need not go into this here because the core claim here is about the truth of our second-order beliefs.

[45] This allows for "Freudian" cases of "repressed" and inaccessible beliefs. – Interestingly, in the *Discours de la Méthode* Descartes – who is sometimes taken as defending the very strong thesis that BB*p* ⌧ B*p* – points out that believing one thing is independent from believing that one does believe that thing; hence people can be ignorant about their own beliefs and they can be wrong about their own beliefs (see René Descartes, *Discours de la Méthode*, in René Descartes, *Oeuvres de Descartes*, eds. Charles Adam and Paul Tannery (Paris: Cerf, 1907-1913), vol. VI, 1-78, especially 23).

be in one of the following three situations:

> I*. occurrently believe at t* that he doesn't believe that *p* and occurrently believe at t* that *p*,
>
> II*. occurrently believe at t* that he doesn't believe that *p* and occurrently believe at t* that not *p*, or
>
> III*. occurrently believe at t* that he doesn't believe that *p* and (occurrently) leave it open at t* whether *p*.

And if the above conjunction principles (Conj-1) and (Conj-2) are plausible, then the following principle would seem plausible, too:

> (Conj-3) If S believes occurrently that he doesn't believe that *p*, is aware of the question whether *p*, and takes a positive or negative stance on *p*, then S has an occurrent belief in the conjunction of the contents of the second-order belief and of his stance on *p*.[46]

As applied to I*, it follows that the subject thinks (something of the form) that

> *p* but I don't believe it.

Given the the alleged parallel to the argument from Moore-paradoxes above, we would have to exclude situation I* as impossible. Only II* and III* would remain, both situations where the subject doesn't believe that *p*.[47] Hence, we have to conclude (in parallel to sections 2-4 above) that B(not B*p*) → not B*p*. And this would get us back into the contradiction above – which would be extremely bad. Given that the argument for *BBp* → *Bp* is strictly parallel, we should also drop the latter.

But this parallel argument for *B(not Bp)* → *not Bp* does not work. Why not? Why should there be such an asymmetry? The crucial point is that "*p* but I don't believe it" (in the sense of "I lack the belief that *p*," not of "I believe that not *p*") is Moore-paradoxical but it doesn't constitute a case of mistrusting one's beliefs. Not

---

[46] The indifference stance would require a different conjunction principle (one parallel to Conj-2):
(Conj-4) If S believes occurrently that he doesn't believe that *p*, is aware of the question whether *p*, and takes an indifference stance on *p*, then S has a conjunctive occurrent belief of the form "I don't believe that *p* and the question whether *p* is unsettled for me."

[47] Leaving something open entails the lack of a belief about the matter. So, III* is a case where S doesn't believe that *p*. One might suspect that case II* has two subcases: one in which S believes occurrently that not *p* without believing occurrently that *p* (II*a), and one in which S believes occurrently that not *p* while also (inconsistently) believing occurrently that *p* (II*b). Doesn't case II*b show that the claim in the text above that in both II* and III* the subject doesn't believe that *p*? No: II*b is ruled out as impossible for the same reasons for which I* is ruled out.

all Moore-paradoxical cases are one's of mistrust towards one's beliefs. Only the commissive cases ("I believe that *p* but not *p*") but not the ommissive ones ("*P* but I don't believe it") are cases of mistrust. One can see the thought that "*p* but I don't believe it" as an admission of epistemic imperfection but not as mistrust of a given belief – there is no belief (that *p*) represented by the subject to itself so that it could be the target of mistrust by the subject. And I don't think it is impossible to think something like "I've won the lottery but I don't believe it" (also think, e.g., about eliminativists about "belief;" see fn.25). This is Moore-paradoxical and irrational but still possible to believe.[48]

5.2. Believing and Holding True

Here is a somewhat lighter problem. Haven't I neglected the difference between believing that *p* and holding-true that *p*? The first entails the second – belief being an attitude of holding true – but not vice versa – as the following shows. Suppose I do not understand some particular thing my friend, the quantum theorist, says, expressing her belief; it is some result in recent quantum physics. She believes that *q* but I do not understand what "q" means. Hence, I cannot believe that *q* (since belief presupposes understanding). But I have good evidence that my friend is speaking sincerely and is usually right about such topics in her field; hence, I have good evidence that what my friend believes about quantum theory is true. Hence, it seems that I might well come to hold the belief (their belief) that *q* true without understanding "*q*" and thus, without having the belief that *q*. I hold-true that *q* but I don't believe it (in the full sense of "believe"). Holding-true is a *de re*-attitude towards the relevant proposition, not a *de dicto*-attitude like belief.

Nothing I have said so far seems to exclude the possibility that someone could falsely believe he understands a sentence and grasps the proposition expressed by it. I might have simply forgotten that I don't understand what "q" means and believe that I do understand it when I don't. In such a case, I would not believe that *q* (because belief presupposes understanding) even if I still hold it true.

[48] To be sure, a thought like that is necessarily false: Given a distribution principle for belief like (Dist), a thinker who holds that "*p* but I don't believe it" also believes *p*; hence, the second conjunct ("I don't believe it") is false and thus also the whole conjunction. – If the subject reflects on her epistemic situation, then she will get from believing that she won the lottery to acknowledging and believing that she so believes. Then it would be very hard to see how she could believe both that she does and does not believe that she has won the lottery. But perhaps one can have beliefs with contradictory contents after all (I will leave this open here). And lacking this step of reflection, the subject could be under the illusion that she has no beliefs about the subject matter. This would be irrational or at least show limited rationality but it would certainly not be impossible.

But (falsely believing that I understand *q*) I might, someone could argue, falsely believe that I believe (and not just hold-true) that *q* when I only hold-true that *q*. In other words, it would seem possible that – contrary to (2) –

(1) S believes that he believes that *p*, but he does not believe that *p*.

My argument above would only have shown that a slightly weaker thesis is true:

(12) If S believes that he believes that *p*, then S holds-true that *p*.

How strong is this objection? First of all, one could simply block this objection early on by insisting that being able to grasp a proposition of the form "B*p*" entails being able to grasp the corresponding proposition *p*. Second, even if it should be possible to believe that one believes something one does in fact not grasp, (12) would still be a very interesting conclusion and almost as strong as (2). Finally, cases of holding-true without belief are exceptions and secondary cases. They are only possible because there are many other cases in which we understand what we hold true. It seems impossible not to have any beliefs and only hold things true; the reason is that holding true involves some belief (e.g., that something is true, etc.). Could there be a subject that holds more propositions true (without believing them) than he believes? What would the life of a subject be like who does not understand the majority or even a substantial portion of what he holds true? Even lacking an argument to the effect that this is impossible, such a scenario seems very unrealistic. So, even if we don't block the objection from the start, we can accept the modification but leave it aside as a secondary case and from now on only look at the standard case of holding a belief true while understanding what one believes.

5.3. Belief and Reflection

Consider the following dialogue (assuming sincerity of the utterances):

February

A: What do you think: How many days does a month have?

B: I believe a month can either have 30 or 31 days!

A: What about February?

B: Oh yes, sorry! I do, of course, not really believe that every month has either 30 or 31 days! Sure, February has less. That's what I really believe!

An objector might point out that such a dialogue makes perfect sense. Doesn't it prove that in his first reply B was wrong about his own beliefs? I do not think so. B did, indeed, believe what he said then. This is compatible with what he

says in his second reply because "really believing" obviously includes something like giving the first-order matter some amount of reflection – which he did not do in his first, spontaneous reply. He believed that months have either 30 or 31 days but he did not "really believe it" in the sense that he did not "believe it after due reflection."[49] This leads to a further clarification of the main thesis: What I have said above concerns the unqualified simple sense of "belief" (concerning *p*) – the sense in which B believed what he said in his first reply. I am not claiming that second-order believing entails reflective first-order believing or that whoever believes that *p*, believes it on the basis of reflection. In other words, this kind of objection does no harm to the thesis I have defended here: $BBp \rightarrow Bp$.[50]

Here is another example which points into the same direction:[51]

Conversion

> Jack was brought up in a very religious family; everyone he knows believes in God and has no doubts about it. Jack then moves to a big city in a different part of the country where he comes into contact with all kinds of people and all kinds of world views. Initially, he is quite shocked but over time gets used to it. In addition, he slowly loses his faith without even noticing it. One day somebody asks him whether he believes in God. Jack replies that he hasn't really thought about if for quite some time but then adds that, sure, he still believes in God. However, after some reflection he denies his first answer: "Sorry, I think I didn't give you a correct answer. I guess I don't believe in God any more."

Prima facie, this seems to be a case in which the person believes that he believes that *p* (that God exists) but does not really believe that *p*. In other words, her second-order belief would be false.

But again, there is a reply like the one to the February-example above. When Jack first answers (sincerely) that he believes in God he does indeed believe in God (given the argument defended here). He might have lost the belief in God before but – given the constitutive nature of second-order beliefs – he "gets it back" (for a very short time) when he acquires (or "reactivates") his second-order belief that he believes in God. As in the February-example above, this belief in God is an ordinary "simple" belief, not a "belief after due reflection (about the subject matter of that first-order belief)" or a "reflective belief" as we might call it. Then,

---

[49] See, for a related distinction, Dan Sperber, "Intuitive and Reflective Beliefs," *Mind and Language* 12 (1997): 67-83.

[50] But doesn't even B's first reply require some reflection? Sure, but this is no objection. What matters is that there is a difference between more or less reflection. We do draw lines between "spontaneous" and "more reflective" responses.

[51] I owe this example to Hilary Kornblith.

after some thinking, he gives up his second-order belief that he believes in God. Since at this time there is no other basis for his belief in God than the second-order belief he is just giving up, the belief in God also goes over board. On top of that, he acquires the belief that he doesn't believe in God. This can be understood in two ways. First, he might simply acquire a second-order belief that he lacks the belief in God - without now believing something different, namely that there is no God. These second-order beliefs (of the form "B not B*p*") need not be true (see section 5.1) but in this case Jack's new second-order belief is true. Second, Jack might in addition acquire a second-order belief that he believes that God doesn't exist. Given our argument here, this entails that he does indeed believe that God doesn't exist. This first-order belief might be a simple, straightforward belief not based on reflection or a belief based on reflection - depending on whether it is only based on his second-order belief but not on reflection about God's existence or whether it is also based on such reflection.

What if Jack already believed that God doesn't exist when he was asked about it? In that case, he was initially not aware of his belief in God's non-existence. When he answered the question and acquired or re-activated his second-order belief that he believes that God exists he also acquired a second belief (a simple, first-order one) about God's existence: namely that He exists. For a short time, until he gave up this belief, he entertained two mutally contradictory beliefs: the belief that God doesn't exist and the belief that God exists. It is a controversial question whether it is possible to believe a contradiction but it is certainly possible to hold two beliefs which contradict each other. In Jack's case he reflected about things and the reflective belief "won" over the simple belief. The inconstency was only short-lived.

5.4. More on Reflection: Epistemic and Semantic

The distinction between beliefs based on reflection and beliefs not so based is quite important here. So, let me add a few more remarks on it. Take Jack's example, again. Was his answer to the question really about God? Jack might have thought along the following lines after his first answer: "Yes, God exists. Oh, wait a minute – what does "God" mean again? Right, now I remember, the creator of the universe who is maximally benevolent, omnisicent, and omnipotent. No, no, no, I don't think that a being like that exists." If this is what is going on when Jack thinks about the question more closely, then reflection is not just about the reasons he might have for a given belief but also about the concepts involved in that belief (the concept of God) or, in other words, about the meanings of the words expressing that belief (the meaning of "God"). The reflection Jack engages in after

his answer might thus either be more of an epistemic nature (only concerned with reasons for a given belief) or more of a semantic nature. Usually, it will be a mixture of both.

One more remark on semantic reflection. Suppose Ernie and Bert are having a chat about math. Bert is asking Ernie whether he believes Goldbach's conjecture is true. Bert has just learned what that is and knows what he is talking about. Ernie first replies, "Sure, haven't you heard about this guy from New Jersey who's proved it?" Then comes the second thought: "Oh, no, wait, that was Fermat's theorem. Gosh, I have no idea. What do you think?" Should we say that Ernie was thinking and talking about Goldbach's conjecture in his first reply? If yes, then we would have to say similar things as in the cases above. If not, then in his first use (in his first reply) of the expression "Goldbach's conjecture" he did not refer to Goldbach's conjecture but rather to, say, Fermat's theorem. In that case, Ernie was not even thinking about and answering the question Bert asked him. In Jack's case semantic reflection led to a fuller understanding of key words whereas in Ernie's case it might have uncovered a simple misunderstanding of core expressions.

One must therefore be quite careful when using examples like *February*, *Conversion* or the Goldbach-example: Insofar as there is simply a change to topic involved between the first and second answer of the subject, nothing at all follows about the possibility of falsely believing one has a certain belief.

5.5. Belief and Behavior

One example that often comes up in discussions about second-order beliefs has to do with psychiatrists:[52]

> Psychiatrists
>
> Suppose Jill does not believe that her parents abandoned her for some time when she was three years old. Her psychiatrist is trying to convince her that she has a "repressed" belief that that was indeed the case. First, Jill rejects the idea: "No, I don't hold that belief." But after the psychiatrist points out some behavioral evidence to the contrary, Jill comes to accept what he says. She acquires the second-order belief that she believes that her parents abandoned her when she was three years old. Isn't this second-order belief just false?

Again, the answer is negative. According to the analysis proposed here, Jill acquires a first-order belief that her parents abandoned her when she was three by accepting the corresponding second-order belief. This is compatible with the fact

[52] Both Susana Nuccetelli and Hilary Kornblith used very similar examples for an objection against my argument.

that she didn't believe that before. The psychiatrist has not only changed her second-order but also her first-order beliefs. What if she not only lacked the relevant belief before her talk to the psychiatrist but, in addition, positively believed that her parents never abandoned her? In that case, we would either have a case of two mutually incompatible and contradictory beliefs or a case in which one belief "wins" over the other and make it disappear (see above). There is nothing problematic with either assumption.

The psychiatrist's case is also interesting because it hints at a difference between two (not the only two!) different ways of acquiring a second-order belief. A person might become convinced by behavioral evidence that she has a certain first-order belief; this kind of evidence is available from a third-person perspective. On the other hand, she might acquire the second-order belief on the basis of reflection about the subject matter of the corresponding first-order belief; this reflection is done from the first-person perspective. In the first case, the resulting first-order belief might rather be a bit more like the acceptance of a theoretical idea whereas in the second case the person might be more wholeheartedly committed to the truth of her first-order belief.[53] This does not entail that one could see one's own beliefs like the beliefs of another person and perhaps even mistrust them. It only means that a person's second-order beliefs can express different kinds of attitudes towards her own beliefs (but always as to her own beliefs). And in both cases, though, the second-order belief entails the first-order belief. [54]

There are more intricacies having to do with this. Consider the case of parachuter P (not necessarily meant as a counter-example):

Parachuting

> P has jumped a couple of times and is convinced that parachuting is not dangerous (and much less risky than driving around in a car which P happily does every day). P is even aware of her belief that parachuting is not dangerous. P is planning to have another jump today. But surprisingly, P just cannot bring herself to jump today. Does this show that P really believed, at least today, that parachuting is dangerous? Does it show that P's second-order belief ("I am not

[53] See L. Jonathan Cohen, *An Essay on Belief and Acceptance* (Oxford: Clarendon, 1992) on this difference.

[54] If the two kinds of attitudes clash, that is, if the person holds from a first-person perspective that she believes that *p* but holds from a third-perspective that she does not believe that *p* (or vice versa), then we have a case of a divided mind if the subject identifies with only "part" of his mind. See fn.30.

> among those who believe that parachuting is dangerous!") was false?[55]

Some propose to distinguish between two kinds of beliefs:[56] avowed beliefs and behavioral beliefs.[57] The first are relatively easily accessible to consciousness but do not necessarily drive our behavior whereas it is the other way around with the second. Should we restrict our thesis that BB$p \rightarrow$ Bp then to avowed beliefs? I don't think we are forced to multiply kinds of beliefs here. It seems more plausible to say that beliefs have many different properties: They represent reality but also drive behavior. The parachuting case can be handled by our approach even if we don't multiply kinds of beliefs. People can be in two minds about things and hold mutually incompatible and contradictory beliefs (that it is dangerous, that it is not dangerous). Apart from that, beliefs might or might not affect behavior and if they do, then their effects can be of quite different kinds (more or less direct, etc.). If they don't, other mental states might drive our behavior: emotions like fear for instance (in case the person is interpreted as not having a belief that jumping is dangerous).[58] What drives our behavior and which of our beliefs lead to action under what conditions, is an empirical question that can only be attacked case by case. Our account is compatible with the parachuting case even if we assume that the behavior of the person reveals a hidden belief that parachuting is dangerous.

5.6. Crimmins and the Idiot

Mark Crimmins has come up with an example that might look like a counter-example to what I am saying here. Here is his paper (I quote in full):

> "'You have known me for years', explained Gonzales, 'But there is something you have not discovered. You know me under two guises, just as Lois Lane knows

[55] See also Shoemaker, "Moore's Paradox" (1996), 89 for such cases.

[56] See Georges Rey, "Toward a Computational Account of *Akrasia* and Self Deception," in *Perspectives on Self Deception*, eds. Amélie O. Rorty and Brian McLaughlin (Berkeley etc.: University of California Press, 1988), 264-296, especially, 272-277; Herbert Fingarette, *Self-Deception* (London: Routledge, 1969), 70, 88.

[57] One could even add a third kind: apart from those beliefs we identify on the basis of behavioral output or on the basis of avowals there would also be those beliefs we identify on the basis of informational input. I will not pursue this here. – An alternative would be to argue for a difference between belief and another kind of state which cannot be assimilated to belief; Tamar Szabó Gendler, "Alief and Belief," *The Journal of Philosophy* 105 (2008): 634-663 introduces the notion of an "alief." Eric Schwitzgebel, "Acting Contrary to Our Professed Beliefs or the Gulf between Judgment and Dispositional Belief," *Pacific Philosophical Quarterly* 91 (2010): 531-553 analyzes such cases as "in-between" cases of belief.

[58] See, e.g.,Neil Levy, "Have I Turned the Stove off? Explaining Everyday Anxiety," *Philosophers' Imprint* 16.2 (2016).

> Superman. You do not realize that I am the person you know under another guise. On that way of thinking about me, you have quite different opinions of me. In fact, you think me an idiot.'
>
> 'Knowing your cleverness,' I replied, 'I must with some embarrassment accept what you say. Since I do not know what guise you mean, I do not know which belief to revise. Until I find out, it seems, I falsely believe that you are an idiot!'"[59]

This is interesting but misleading in a subtle way. Crimmins believes something like this:

> Gonzales is no idiot.

Crimmins also learns this:

> ∃x (x=Gonzales & I believe of x that he is an idiot).

The only thing of relevance here Crimmins can infer is:

> I falsely believe *of* Gonzales that he is an idiot,

or, in other words:

> ∃x (x=Gonzales & I falsely believe of x that he is an idiot).

However, Crimmins cannot infer:

> I falsely believe *that* Gonzales is an idiot.

Crimmins can only ascribe a certain *de re* belief to himself but not the relevant *de dicto* belief. Since we are only dealing with *de dicto* beliefs here, Crimmins's case does not constitute a counter-example.[60]

So much for some objections to my main claim. It turns out, I think, that they do not work against our constitutive view of knowing one's beliefs.

6. Conclusion

I have argued for the claim that $BBp \rightarrow Bp$: If one believes that one believes that *p*, then one believes that *p*. If Mary believes that she believes that justice is the highest virtue, then she does indeed believe that justice is the highest virtue. This is a surprising claim: Sometimes "believing makes it so." It goes against what many people, especially philosophers, psychologists and cognitive scientists, believe. It might seem even more surprising that there are good arguments supporting this

[59] Mark Crimmins, "I Falsely Believe that *P*," *Analysis* 52 (1992): 191.

[60] See also Alan Hajek, Daniel Stoljar, "Crimmins, Gonzales and Moore," *Analysis* 61 (2001): 208-213, and David M. Rosenthal, "Moore's Paradox and Crimmins's Case," *Analysis* 62 (2002): 167-171; Williams, "Wittgenstein," sec.10 is very close to what I am saying here.

constitutive account of knowledge of one's own beliefs. I have used an argument which is based on considerations on Moore's paradox and on the impossibility of mistrusting one's own beliefs. Accepting the claim that BB$p \rightarrow$ Bp certainly has farreaching consequences for the way we should think about self-knowledge. [61]

[61] I would like to thank Sven Bernecker, Monika Betzler, Vivienne Brown, Gisela Cramer, Richard Eldridge, David Hemp, Hilary Kornblith, Teresa Marques, Susana Nuccetelli, Neil Roughley, John Williams, Truls Wyller und audiences in Maribor, Seattle, Belfast, and Aberdeen as well as some referees for comments.