One Variety of Self-Knowledge: Constitutivism as Constructivism*

Annalisa Coliva
University of Modena and Reggio Emilia & COGITO

It is a commonplace that we know our own minds, viz. that we know our own sensations, feelings, perceptions, imaginations, emotions as well as propositional attitudes, such as beliefs, desires, intentions, hopes, wishes and so on. Still, there is little consensus over what would count as a sound philosophical explanation of our knowledge of each of these kinds of mental states. For, on the one hand, a variety of competing encompassing theories of self-knowledge are available nowadays.[1] On the other, given the intrinsic differences among the various kinds of mental states we can enjoy, it may well be that the most apt attitude towards self-knowledge should in fact be *pluralistic*—that is, such as to allow for different accounts of how we know each of these various kinds of mental states.

* I would like to thank people in attendance at various presentations of previous versions of this paper such as the participants to "The Self and Self-Knowledge I" conference, held in Bigorio (CH), and kindly supported by the Fonds nationals Suisse (Project number 1115-05572): Dorit Bar-On, Akeel Bilgrami, Jane Heal, Martine Nida-Rümelin, Lucy O'Brien, Eva Picardi, Jim Pryor, Carol Rovane, Gianfranco Soldati, Crispin Wright. My thanks also to Achille Varzi for helpful discussions, to Paulo Faria, Elisabeth Pacherie, Joelle Proust, Barry Smith, Isidora Stojanovic and to the rest of the audience at Institut Jean Nicod, Paris, as well as to people in attendance at King's College London, in particular Charles Travis and Keith Hossack. I would also like to thank the audience at the ECAP conference in Lisbon 2005, at the Italo-Spanish workshop in Bologna in 2005 and at the SIFA conference in Genova in 2004. I also take this opportunity to express my gratitude to my colleagues and friends at COGITO Research Centre, in particular, Paolo Leonardi, Eva Picardi, Sebastiano Moruzzi, Giorgio Volpe, Delia Belleri, Michele Palmira, as well as Carla Bagnoli and Mario Alai for helpful discussion on the penultimate draft of this paper. Most of all, however, my thanks go to the Italian Academy at Columbia University (NY) and the Alexander von Humboldt Stiftung for generous support and to my host in Heidelberg, Professor Andreas Kemmerling, for providing ideal settings in which much of the research that eventually led to the present paper was conducted.

[1] See, for instance, Bar-On, D. 2004 *Speaking My Mind. Expression and Self-Knowledge*, Oxford, Oxford University Press. Another potentially encompassing account of self-knowledge is the one presented in Peacocke, C. 1999 *Being Known*, Oxford, Clarendon Press, Ch. 5. See also his 2004 *The Realm of Reason*, Oxford, Oxford University Press, Chs. 3, 8 for an account of self-knowledge of perceptions and an account of the emotions, which would be compatible with Peacocke's other pronouncements on self-knowledge.

Sydney Shoemaker[2] and Crispin Wright[3] have been among the first theorists to propose a so-called "constitutive" account of self-knowledge. Constitutive accounts are designed to apply mainly to knowledge of our own propositional attitudes. For it is a tenet of this kind of account that having knowledge of one's first order mental states is (at least) a necessary condition for having those mental states. Clearly, this would make little sense for sensations and perceptions, as well as for some kind of emotions, which we may want to grant to creatures such as infants and higher-order mammals whom, however, we think would lack knowledge of their own mental states. This, therefore, invites the idea that self-knowledge as a whole should be explained in a variety of ways.[4]

Despite Akeel Bilgrami's new vigorous and thought-provoking attempt at defending a constitutive account of our knowledge of propositional attitudes,[5] constitutive accounts have recently come under attack.[6] For many theorists working in this area are getting increasingly uneasy with the idea, which has by now become the central tenet of constitutive accounts, that the immediate and authoritative way in which each of us knows her own propositional attitudes is *not* the result of any cognitive achievement, consisting, rather, in a pair of two conceptual truths which can be variously redeemed. In fact, many theorists are now trying to defend the idea that although self-knowledge of propositional attitudes is neither observational, nor inferential, just as constitutivism holds, it nevertheless counts as a genuine kind of knowledge: for it would consist in having true beliefs about one's own first-order propositional attitudes, which would exist independently of one's knowledge of them, for the *reason* that one has them. So, self-knowledge,

[2] See also Shoemaker, S. 1968 "Self-reference and self-awareness", *Journal of Philosophy* 65, pp. 555-578; Shoemaker, S. 1986 "Introspection and the self", *Midwest Studies in Philosophy* 10. Reprinted in Shoemaker, S. 1996a *The First Person Perspective and Other Essays*, Cambridge-New York, Cambridge University Press, pp. 3-24; Shoemaker, S. 1988 "On knowing one's mind", *Philosophical Perspectives* 2. Reprinted in Shoemaker, S. 1996a, pp. 25-49; Shoemaker, S. 1990 "First person access", *Philosophical Perspectives* 4. Reprinted in Shoemaker, S. 1996a, pp. 50-73; Shoemaker, S. 1996b "Self-knowledge and inner sense. Lectures I-III", in Shoemaker, S. 1996a, pp. 201-268. Heal, J. 2002 "First person authority", *Proceedings of the Aristotelian Society* 102. Reprinted in Heal, J. 2003 *Mind, Reason and Imagination*, Cambridge, Cambridge University Press, pp. 273-288.

[3] Wright, C. 1989a "Wittgenstein's rule-following considerations and the central project of theoretical linguistics", in A. George (ed.) *Reflections on Chomsky*, Blackwell. Reprinted in Wright, C. 2001a *Rails to Infinity*, Cambridge (Mass), Harvard University Press, pp. 170-213; Wright, C. 1989b "Wittgenstein's later philosophy of mind: sensation, privacy and intentions", *Journal of Philosophy* 86, pp. 222-234. Reprinted in Wright, C. 2001a, pp. 291-318; Wright, C. 1998 "Self-knowledge: the Wittgensteinian legacy", in Wright, C., Smith, B., Macdonald, C. (eds.) *Knowing Our Own Minds*, Oxford, Clarendon Press, pp. 13-45; Wright, C. 2001b "The problem of self-knowledge (I) and (II)", in Wright, C. 2001a, pp. 319-373.

[4] As Patrizia Pedrini kindly pointed out to me, recently Matthew Boyle has argued for this very conclusion. See his 2009 "Two kinds of self-knowledge", *Philosophy and Phenomenological Research* 77/1, pp. 133-164.

[5] Bilgrami, A. 2006 *Self-Knowledge and Resentment*, Cambridge (Mass.), Harvard University Press. Bilgrami's account allows him to connect self-knowledge, agency, value and intentionality, which he sees as fundamentally integrated.

[6] See Peacocke 1999, 2004; Moran, R. 2001 *Authority and Estrangement*, Princeton, Princeton University Press; Bar-On 2004, Ch. 9 which, interestingly aims to combine this view with expressivism; O'Brien, L. 2007 *Self-Knowing Agents*, Oxford, Oxford University Press. Much of the inspiration for these accounts comes from Burge, T. 1996 "Our entitlement to self-knowledge", *Proceedings of the Aristotelian Society* 96, pp. 1-26.

despite its being neither observational nor inferential, would not be *baseless*,[7] or *groundless*.[8] For one's second-order (true) beliefs would in fact be based on, or grounded in the corresponding first-order mental states and held because one is somehow aware of them.[9]

An initial difficulty with these latter accounts is that it is quite hard to see how *one's own available reason* or *warrant*[10] for one's psychological self-ascriptions could be anything else but the avowal itself. For, if asked "How do you know that—say—you believe that P?" I could only answer by repeating my avowal, viz. by saying "Because I do believe that P". Hence, it is not clear how subjects could provide independent reasons for their avowals. More generally, as I have argued elsewhere,[11] if, in order for first-order mental states to function as reasons of the corresponding self-ascriptions, subjects need to be aware of them, depending on what notion of awareness one favours —propositional or phenomenal—, it may well turn out either that self-knowledge of the basing state be presupposed; or else that awareness of it falls short of providing one with a genuine reason for the self-ascription. Thus, I think constitutive accounts remain the most promising way of looking at our knowledge of our own propositional attitudes. In this paper, I will present a new brand of constitutivism which, while finding its inspiration in the works of Wright and Bilgrami, seems to me to have the resources to do better than its predecessors on a number of fronts.

1. The constitutive thesis

According to constitutive theorists, self-knowledge can be based neither on observation nor on inference for either would fail to account for two features of it which seem intuitively compelling. Namely: so-called "transparency" and "authority". Transparency amounts to the idea that at least in a large class of cases the occurrence of one's own mental states is of a piece with one's awareness of their kind and content. If, for instance, I were asked "What are you thinking of right now?" I would be able to answer immediately, without conducting any inquiry. Whatever kind of thought is now crossing my mind, it would seem to be immediately known to me, both with respect to its kind —its being a belief, a desire, a wish, etc.—and its content. It seems to make no sense to suppose that I should somehow find out what I am now thinking of by observing myself and my behaviour. Nor would it make sense to suppose that such knowledge would in fact depend on observing mental states somehow luminously presented in my mental arena. After all, mental states are not kinds of

[7] The expression is McDowell's, although not used in the context of providing a positive account of self-knowledge. McDowell, J. 1998 "Response to Crispin Wright", in C. Wright, B. Smith and C. Macdonald (eds.), pp. 47-62, at p. 48 and *infra*.

[8] This is Wright's way of characterizing self-knowledge in his works. Cf. fn. 3.

[9] Notice that what would be baseless or groundless is the self-ascription of the relevant mental states, according to the constitutive account, not the first-order mental states themselves which may indeed be based on grounds and evidence, as we shall see at much greater length in the following.

[10] All these accounts are advertised as internalist.

[11] For a detailed criticism of Peacocke's position, see Coliva 2008 "Peacocke's self-knowledge", *Ratio* 21/1, pp. 13-27. Some qualms are raised also in Bilgrami 2006, pp. 134-139. See also the essays by Heal and Soldati in this volume.

objects we may observe, and to think otherwise would in fact depend on holding on to a Cartesian conception of the mind, whose limits and intrinsic incoherence have been variously exposed.[12] Authority, in contrast, consists in the fact that, at least in the vast majority of cases, one's sincere and competent avowals can't rationally be challenged. It makes no sense, at least in most cases, as we shall review in the following, to challenge a subject who sincerely avows "I believe that summers in Greece are really too hot" and is competent with respect to the relevant concepts, by saying "How do know that you believe it?—Give me your grounds for your claim", and so on.[13]

Furthermore, transparency and authority seem to be *a priori* and necessary features of our knowledge of our own mental states, not just mere contingencies.[14] For there is at least a general presumption that people will know their own mental states and that their avowals will be correct. We may think of exceptions both to transparency and authority, such as unconscious mental states, which will be there but won't be self-known, or cases of self-deception, where a subject might think she has a given mental state she doesn't really seem to have. But either exception seems peculiar. Not knowing that one has a serious hatred towards all other male subjects as the result of one's Oedipus complex is not quite like ignoring what one is thinking of right now. The first kind of ignorance would not impair the idea that we are dealing with a human being capable of a real mental life. The latter, in contrast,[15] would: it seems to be part and parcel of our own conception of adult human beings' mentality that they have knowledge of their own occurrent mental states. Furthermore, supposing that subjects could routinely self-ascribe beliefs and desires they don't

[12] See, for instance, Wright 2001b, pp. 331-340; Bar-On 2004, Ch. 2, pp. 37-46; Bilgrami 2006, pp. 3-8. Most criticisms rely on Wittgenstein's considerations against the very intelligibility of a private language. A different kind of criticism can be found in Shoemaker 1996, Lecture I in particular.

[13] Obviously what wouldn't make sense is to challenge a subject's psychological self-ascription, not her grounds for holding that summers in Greece are really too hot (cf. fn. 9). Furthermore, as the qualifications suggest, it wouldn't make sense provided there were no reasons to think that such a subject may be self-deceived. There will be more on self-deception in the following.

[14] The fact that transparency and authority are held to be a priori and necessary features of self-knowledge as opposed to mere contingencies sets constitutivists apart from other theorists, such as functionalists, who may think that it is part of the functional role of propositional attitudes that they give rise to a correct second-order belief about them. Indeed, David Armstrong 1968 *A Materialist Theory of the Mind*, London, Routledge, though no functionalist at all, is prominent for holding such a view. Notice, however, that on his picture failures at self-knowledge, either for lack of transparency or of authority, would be due to the malfunctioning of a subpersonal cognitive mechanism, which either does not produce second-order beliefs, or gives rise to erroneous ones. It should be noted, however, that such failures would not impair a subject's rationality—not any more than being colour-blind should impair one's rationality at using colour concepts. Constitutivists, in contrast, maintain that failing to know one's own mental states or being massively mistaken with respect to them would rightly make us suspicious of dealing with a subject who is rational and in possession of the relevant conceptual repertoire. On this issue, see, for instance, Bar-On 2004, pp. 95-104, Bilgrami 2006, ch. 1, Wright 1998, p. 17. I discuss and criticise Armstrong's position in more detail in Coliva, A. 2006 "Self-knowledge: another constitutive view", *Preprint Dipartimento di Filosofia Università di Bologna* 28, pp. 101-121 (esp. at pp. 104-106). True, as an anonymous referee has pointed out, there may be other accounts of self-knowledge, beside constitutive ones, which exclude massive failures of self-knowledge, if subjects are to be granted with the relevant conceptual repertoire. These proposals would have to be discussed on merit and, in particular, it should be seen whether they would be compatible with the claim—which seems to me distinctive of constitutivism—that radical failures of self-knowledge would impair a subject's rationality. In any event, the considerations proposed in the main text are merely meant to make an at least prima facie case for the plausibility and interest of constitutive positions.

[15] Under certain constraints I will indicate in the following.

really have would hinder the idea that they possess the relevant psychological concepts, and, arguably, that we are dealing with rational subjects at all.

So bar these peculiar (and perhaps only seeming) exceptions, there seems to be a general presumption that mental states are transparent to the subjects who have them and that when they avow their own mental states they are correct. To stress, transparency and authority don't seem to be just empirical generalisations. Rather they seem to be of a piece with our conception of ourselves and others as endowed with a real and normal mental life, as rational agents, and, finally, of a piece with our linguistic practice of making avowals, in which our conceptual mastery is deployed.

Yet, no bit of knowledge based on observation is either "transparent" or "authoritative". For, by definition, it will always be based on a however minimal empirical inquiry, and will remain open to rational challenges, at least in principle. Similarly, our inferential knowledge will always be based on connecting our observations with some kind of theory, thus failing to be transparent, and, obviously, it will always be amenable to rational scrutiny. Since we don't have any other way of knowing empirical truths, other than by observation or by inference, we should conclude that self-knowledge is in fact based on *nothing*.[16] To repeat, what this means is that so-called "self-knowledge" is not a kind of cognitive achievement after all, consisting in holding a true belief on the basis of having reasons for it—no matter how you might construe these reasons. Therefore, if knowledge is understood in such a usual way, it is somehow a misnomer to call it "knowledge". Rather, what we call "self-knowledge"—that is the distinctive kind of authority we recognise to our fellow humans (and to ourselves) over their own mental states, as well as the distinctively immediate, or transparent way in which they are aware of them—are guaranteed to hold *a priori*.

In more detail, all constitutive theorists agree that a suitably qualified version of the following thesis holds *a priori* and that, in fact, it is true as a matter of *conceptual necessity*.

> **Constitutive Thesis**: given certain conditions C, S believes/desires/intends/wishes/hopes that P if and only if S believes (or judges) that she believes/desires/intends/wishes/hopes that P.[17]

However, constitutive theorists debate the following:

(i) what the *grounds* of the constitutive thesis are—e.g. is it grounded in the linguistic practice of making psychological avowals (Wright); or in the notion of rationality

[16] See Wright, C. 1989b, p. 312 and Boghossian, P. 1989 "Content and self-knowledge", *Philosophical Topics* 17, pp. 5-26 (at p. 5).

[17] I think that the formulation of the second half of the constitutive thesis is irrelevant, as long as one holds that judgment brings about belief. This view can be found, for instance, in Peacocke 1999, p. 238, as well as in Scanlon, T. 1998 *What We Owe to Each Other*, Harvard, Harvard University Press, Ch. 1; but also in Moran 2001, p. 116.

(Shoemaker); or, else, in the notion of deliberative agency (Bilgrami)? Depending on their answers to such questions, they will return different characterizations of the C-conditions which are supposed to constrain the constitutive thesis. Furthermore, what they debate is

(ii) how to interpret the thesis and, in particular, what kind of *metaphysical implications* it has. On some constitutive accounts, to judge that, say, one believes that P does (at least partially and in some cases) bring about the corresponding first-order propositional attitude (Shoemaker, Wright, and myself, as we shall see). Hence, enjoying the latter is actually *constituted* by one's believing to be in such a state. On some other, weaker constitutive views, such as Bilgrami's, the relevant self-ascriptions are not intended as bringing about the corresponding first-order propositional attitudes. Hence, the constitutive thesis is read as merely entailing that self-knowledge is both a necessary and a sufficient condition for the corresponding first-order mental states.[18]

Here, however, I won't look at the details of other constitutive accounts and will simply let my (dis)agreement emerge in the course of the presentation of my own positive proposal.

2. Transparency

Let us focus on the left-to-right side of the constitutive thesis, which elevates transparency to the rank of a conceptual truth. Clearly, we should specify the C-conditions so as to impose the obvious constraint that the biconditional should hold only for a lucid and sincere subject. Even so, however, it remains that maintaining it is a conceptual truth that if such a subject has a given first-order belief/desire/intention/wish/hope and so on she will believe she does, seems to be bound to generate some critical reactions. Two are most likely.

As we have already anticipated, there are unconscious mental states. If one allows for them,[19] then they would be there even if one is in no position to self-ascribe them. Furthermore, we are now

[18] Sometimes people worry about the fact that such a position could really qualify as constitutive (O'Brien in conversation). Insofar as constitutivism is taken to be individuated by its adherence to the view that self-knowledge is not the result of any cognitive achievement and that it consists in maintaining that the constitutive thesis holds as a matter of conceptual necessity, I think we can embrace Bilgrami's position within the scope of constitutivism. So, we could distinguish between "weak" and "strong" forms of constitutivism: they would all hold that self-knowledge isn't the result of any cognitive achievement, but only strong ones would add that first-order mental states are (at least partially and in some cases) constituted by having the corresponding second-order ones. Notice, moreover, that weak forms of constitutivism would be applicable also to phenomenal self-knowledge (as long as we were dealing with creatures endowed with the relevant conceptual repertoire and we characterised the C-conditions accordingly, with obviously no reference to propositional attitudes as commitments). By contrast, I don't think strong forms of constitutivism could sensibly carry over to our knowledge of our own sensations and other non-propositional mental states. For the fact that we share them with infants and at least higher-order mammals seems to me incompatible with the right-to-left side of the Constitutive Thesis. That is to say, there seems to be no scope for the view that these mental states could be at least partly constituted by one's own judgment that one has them.

[19] Personally I am not skeptical with respect to them. But, in case one were so skeptical, one possible counterexample to transparency would disappear.

conversant with the practice of ascribing at least beliefs and desires to higher-order mammals and infants to explain their purposive behaviour that can't be explained simply in a causal-nomological manner.[20] Still, we don't want to say that these creatures have knowledge of their own mental states. So, this would be another case where there would be first-order propositional attitudes but no second-order beliefs about them.

2.1 Mental states as commitments

In order to answer these objections I think it is useful to point out that, on reflection and contrary to what mainstream philosophy of mind seems to hold, our notion of an intentional mental state isn't univocal. On the one hand, there are intentional mental states that we might call "mental states *as dispositions*" or "*non-judgement-sensitive* mental states".[21] Admittedly, this would be a very heterogeneous class, whose width may be hard to determine exactly. However, what will characterise the mental states that belong to it is at least the following:

(a) these mental states aren't the result of a conscious deliberation, i.e. a judgment, on a subject's part, based on considering and, in particular, on *assessing* evidence in favour of P (or of P is worth pursuing, it would be good if P happened, etc.);

(b) these mental states aren't within one's direct control, being rather something one finds oneself saddled with;

(c) hence, these mental states aren't something one will be held rationally responsible for.

Some examples of mental states that will satisfy these conditions are (i) mental states that aren't formed by being able to *assess* evidence in favour of P (in the case of beliefs) or of P would be good to have (in the case of desires, intentions and hopes); though one may form them in response to available evidence in favour of P, or of P would be good to have, if presented with it.[22] (ii) Mental states that are attributed to subjects to make sense of their behaviour, of which they themselves may be entirely *ignorant*. (iii) The latter class of mental states may comprise, but isn't exhausted by, *unconscious* mental states of a Freudian kind, which, however, can be operative in shaping a subject's behaviour. (iv) Also mental states that are self-attributed on the basis of an act of *self-interpretation*, by finding them out through the observation of one's own behaviour and other immediately self-known mental states, will fall into this category. For self-interpretation, when successful, makes one aware of a mental state that is already there, yet isn't "one's making",

[20] Again, someone might deny the legitimacy of such a practice. Although I think we shouldn't be too cavalier in ascribing beliefs and desires to these creatures, I think that qualified ascriptions of beliefs and desires to them are fine. In the following I will be more precise about the form these qualifications should take.

[21] Bilgrami 2006, Ch. 5 in particular, distinguishes between mental states as "dispositions" and as "commitments". Scanlon and Moran between "brute" or "non-judgment sensitive" and "judgment-sensitive" mental states.

[22] Brute urges and needs would fall into this category; but also, I think, those dispositional mental states we may ascribe to a-conceptual creatures to make sense of their intelligent behaviour, which, while responsive to some evidence, aren't dependent on its appraisal.

but, rather, something one finds oneself saddled with.[23] Finally, and as the converse of self-interpretation, (v) there may be mental states, which one can inferentially *predict* will assail one, in given circumstances, which, however, *won't be within one's direct control*.[24]

Yet, manifestly, adult human beings also have different kinds of mental states, namely, mental states that depend on a judgment based on the *assessment* of the evidence at subjects' disposal, and that, for this reason, are within their control and for which they are held rationally responsible. Call them "intentional mental states as *commitments*", or "*judgement-sensitive* mental states".[25] Although the word "commitment" may have become common currency in philosophical literature nowadays,[26] I think there is still little agreement among its users about its meaning. In my view, what is essential to commitments such as to make them, in effect, very close to "judgement-sensitive" beliefs, desires, intentions, wishes, hopes and so on, is the following:

(a) that they are the result of an action—the mental action of *judging* that P is the case (or worth pursuing/having)—on the subject's part, on the basis of considering and hence of *assessing* evidence for P (is worth pursuing/having);[27]

(b) that these mental states are *normatively constrained*, i.e. they must respond to the principles governing theoretical and practical reasoning;

(c) and, in particular, they are so constrained (also) *from the subject's own point of view*;

(d) that they are mental states for which the subject is held *rationally responsible*.[28]

---

[23] A nice example, although not a case of propositional attitude, is provided by Jane Austen in her novel *Emma*, when the protagonist finds out about her love for Mr. Knightly, her long-lasting friend, by reflection and inference on her own immediately available feelings of jealousy at the prospect that Mr. Knightly could return another woman's feelings. The example is presented and discussed in Wright 1998, pp. 15-16, borrowed from Tanney, J. 1996 "A constructivist picture of self-knowledge", *Philosophy* 71, pp. 405-422. (Notice, however, that Tanney is concerned with the kind of construction of one's own mind, which could occur in self-interpretation. Here, in contrast, I will be concerned with constructivism regarding one's own immediate avowals of propositional attitudes). Analogous examples could easily be construed for the case of propositional attitudes. Giorgio Volpe has kindly pointed out to me that also Schopenhauer in *On Freedom of the Human Will* holds the view that a person's character traits are known to him through reflection and inference on his past behavior.

[24] Perhaps, due to one's long-lasting self-observations, one will know that if one were to work in an unsupportive environment for a while, one would start losing one's self-confidence and believing that one's work is meaningless, or of poor quality. The characteristic feature of these mental states is that one would seem to find oneself *saddled with them*, even if one were rationally able to find reasons which should make one think differently.

[25] Bilgrami 2006, p. 213; Scanlon 1998, Ch. 1 and Moran 2001, p. 116.

[26] Bilgrami makes extensive use of the term; Robert Brandom too, although he is more interested in stressing the social dimension of commitments, than the former (or indeed myself). Furthermore, it is not my contention, somehow built in in the very notion of a commitment, that one should have knowledge of all the logical consequences of one's own beliefs and further propositional attitudes. As Bilgrami points out (2006, pp. 371-2, fn. 7, but see also pp. 376-377, fn. 20), the origin of the use of this term to refer to intentional states (or at least to a class of them) goes back to Isaac Levi.

[27] This is the main difference between my account of commitments and Bilgrami's. For, on his view, commitments aren't dependent on a subject's judgment.

[28] This is the constraint Bilgrami identifies as essential to commitments, from which, on his view, (b) and (c) follow. However, he gives a moral or evaluative twist to it I would resist. For, on his view, not only would one be held *rationally* responsible for one's commitments, but also *accountable* at large. For instance, one might be *reproached* or *resented* for having certain commitments (cf. Bilgrami 2006, p. 226). To my mind, however, specified in the way Bilgrami characterises it, (d) is not sufficient to mark out the contrast between commitments and dispositions, because one can criticise or be criticised, and accept to be criticised, also (for) one's own dispositions, such as the disposition to smoke, or, to take a more loaded example, for wanting to get rid of other male opponents as a result of an unresolved

So, in my view, mental states as commitments depend on a subject's *deliberation* with respect to P, in the case of belief, of "P ought to be pursued" and of "It would be good (for me) if P were the case" in the case of desires and intentions, hopes and wishes, and so on, based on considering and evaluating evidence for P or for "P ought to be pursued", etc.[29] Furthermore, should countervailing evidence come in, then a subject *ought* to withdraw from holding P, "P ought to be pursued" and "It would be good (for me) if P were the case" etc., thus, withdraw from one's belief that P is the case or from one's desire/intention/hope/wish that P should obtain. Finally, precisely because such "oughts" would have to be appreciated by the subject herself, were she not to withdraw from her beliefs and further propositional attitudes as commitments in light of counter-evidence, she would not only incur *rational criticism*, but she should also *accept* to incur it, since she wouldn't have—the phrase comes in handy—"lived up to her commitments".

Now, some clarifications are in order. First, to say that beliefs (as well as other propositional attitudes) as commitments are the result of a deliberation and fall within one's responsibility doesn't "involve us in any sort of voluntarism about [their] formation […], any more than we need to see ordinary argument with others as aiming at getting one's interlocutor somehow to adopt a new belief by sheer act of arbitrary will".[30] That is to say, forming certain mental states by considering relevant pieces of evidence is a rational action, yet not an act of arbitrary will. Indeed, being aware of evidence in favour of P, and being unaware of countervailing evidence, ought to give rise to one's judgment that P is the case, and thus to the corresponding belief (similar considerations would hold for other propositional attitudes and the kind of evidence with respect to which they are the rational response). Hence, it would be a sign of irrationality not to form that belief, given favourable evidence for P.[31] If one did not form such a belief, one should incur and ought to accept to incur

Oedipus complex. But, surely, neither mental state is the result of a subject's action, for which one can be held rationally responsible, although one may be considered "badly"—in Bilgrami's extended sense of the term—for having it. It is then not by chance that, as a matter of fact, Bilgrami ends up endorsing the view that "we do have transparent self-knowledge of mental dispositions" (Bilgrami 2006, p. 287). I find this conclusion unpalatable, for, surely, when we do get knowledge of our unconscious mental states we obtain it through a process of self-interpretation or of analysis (that may or may not be guided by a therapist) relevantly similar to the ways in which we may come to attribute mental states to others. So, it seems to me that whatever knowledge we may eventually gain of our unconscious mental states it is not transparent, and is actually grounded in observation and inference. Moreover, in the case of dispositions such as the disposition to smoke, we certainly don't want to account for their transparency by mobilising something like the constitutive thesis, read in a strong sense (see fn. 18). For that would have the unattractive consequence of making one's own dispositions the result (at least partly) of one's beliefs about them. So, in the case of dispositions such as the disposition to smoke a cigarette, when they are known immediately—that is, when the corresponding *urge* manifests itself—I would rather account for that knowledge along purely expressivist lines. The kind of more general knowledge of one's own generic disposition to smoke, in contrast, would be self-known through a process of self-observation and interpretation; hence, in a third-personal way.

[29] This is a sketchy account that, however, doesn't prevent obvious forms of local holism between mental states from arising. Indeed, viewing also desires, intentions and other propositional attitudes, beside beliefs, as commitments, tightly connects them with *believing* that their contents are worth-pursuing or would be good for one if actualised. This, in my view, is a plus and not a deficiency of the present proposal.

[30] Moran 2001, p. 120.

[31] I said that it would be a sign of irrationality not to form certain mental states as commitments, given certain pieces of evidence, and this might invite the objection that, at least in the case of intentions, one could fail to form them, upon

criticism. Conversely, forming the belief that P when no evidence in favour of P is at one's disposal, let alone contrary, undefeated evidence is available to one, would not be the rational thing for one to do. If one did, then one should accept and ought to agree to accept criticism.

Secondly, it is important to emphasise that a subject may not always arrive at her beliefs and further propositional attitudes as a result of *conscious* consideration of the evidence for P (or else, for P ought to be pursued, or for "it would be good (for me) if P were the case"). All the present account of commitments requires is that she must be able to offer her evidence in support of her beliefs and judgement-sensitive desires and other propositional attitudes if asked to give her grounds for them—that is, the reasons why she holds them, not the reasons why she *believes* she does.[32] This is part of what distinguishes commitments from dispositions, which while perhaps based on evidence, aren't based on the assessment—not even the potential assessment—of such evidence.

Thirdly, it may sound surprising that also desires could be seen as brought about by rational deliberation. But I think it is important to keep in mind that here I am not concerned with what we might call "brute" desires such as lust or hunger, but only with rationally held ones.[33] For instance, one may rationally desire to provide one's child with the best possible education. If one does so, it will be for reasons, which, as such, may be further assessed. Were it to turn out—quite implausibly—that countervailing considerations should outweigh that desire, one should withhold from it, if rational.

Finally, it has to be registered that although the distinction between mental states as commitments and as dispositions may be misleading, insofar as it might suggest the idea that commitments *exclude* having behavioural dispositions (at large),[34] on my understanding of them,

---

having suitable evidence in their favour, as a result of mere weakness of will. My view is that weakness of will could only prevent one from *acting* on the basis of a given intention, but would not impinge on its formation. One might then suggest that despite having evidence in favour of "It would be good (for me) if P were the case" a (lucid and attentive) subject could fail to form the corresponding intention because of other considerations. Hence, rationality doesn't require forming the relevant intentions upon having at one's disposal certain pieces of evidence. I think that is all right, but that it merely shows that the subject didn't have *sufficient* reasons to form a given intention. Hence, it is obvious, though I haven't so far added nor will in the following add such a qualification, that when I claim that there are intentions as commitments that are based on evidence, I am in fact talking of those intentions we form on the basis of *sufficient* reasons for them. Again (cf. fn. 29), this shows that, inevitably, there will be forms of local holism in the formation of one's mental states, which is all to be expected.

[32] See also Moran 2001, p. 116.

[33] See also Scanlon 1998; Moran 2001, p. 116 and Bilgrami 2006, p. 214. These "oughts" are what distinguishes beliefs and desires as commitments from mere drives and brute dispositions: if I believe/desire that P as a commitment, then I *ought* to do so on the basis of *evidence* and ought to *withhold* from it in case contrary evidence came up. Obviously no such "oughts" hold for drives, e.g. the urge to smoke a cigarette after dinner, which will persist no matter what amount of counterevidence will be considered against the advisability of such a practice; or brute dispositions, such as the disposition to form a certain thought upon hearing a given word, tune, etc. As already remarked (see fn. 28), it may well be that the best account of our knowledge of our "brute" desires, which are not known through self-interpretation, should be given along expressivist lines.

[34] I am not sure Bilgrami really avoids this risk (see for instance, 2006, pp. 210, 226), although, on his view, commitments consist in having the (second-order) disposition to be self-critical or to accept criticism from another person, if one fails to live up to one's commitments (see Bilgrami 2006, p. 226). Bilgrami seems to see commitments

mental states as commitments will in fact *have to* be accompanied by the relevant set of dispositions:[35] for instance, if I believe that P as a commitment then I ought to be disposed (*ceteris paribus*) to use P as a premise in a piece of practical or theoretical reasoning, to assert "P", to give grounds for my claim if challenged, as well as to withdraw from it in case contrary evidence came up. Similarly, in the case of a desire as a commitment that P, I ought to be disposed (*ceteris paribus*) to seek means to make P happen, or withdraw from it if it were shown that it is not worth-pursuing, and so on.[36] If I didn't actualise such dispositions, I should then be prepared to accept criticism for having failed to live up to my commitments.

Now, let us go back to the apparent counterexamples. First of all, it seems safe to hold that while infants and higher-order mammals may of course have "brute" desires and be able to react and interact in intelligent ways with their environment, thus (perhaps) justifying the fact that we may want to attribute beliefs and desires to them, they certainly don't have mental states as a result of judgment and of actively bringing evidence and practical considerations to bear on what they believe and desire, by considering and assessing such evidence and practical considerations, not even potentially. Hence, whatever kind of mental states they can actually enjoy, they can't have mental states as commitments.[37] Similarly, unconscious mental states aren't, obviously, brought about by judgments sensitive to epistemic evidence and practical considerations. They are produced, rather, by experiences we have had, mostly in early years, but aren't formed by consciously assenting to certain contents in light of assessed evidence for them. Hence, they aren't commitments. Thus, infants' and animals' mental states as well as unconscious ones won't be counterexamples to the left-to-right side of the constitutive thesis, once that thesis is appropriately qualified. That is to say, if it is taken to hold *only* for mental states as *commitments*. Yet, we must

---

and (first-order) dispositions as mutually exclusive types of mental states mostly because of his preoccupation to avoid naturalism. A discussion of this topic will have to be postponed to another occasion (but see fn. 37 for a bit more detail).

[35] Here is another difference between Bilgrami's understanding of commitments and mine. For, on his view, although commitments *may* be accompanied by dispositions they *need not* be. See Bilgrami 2006, pp. 214-215; 225. So, on his account, one could have a commitment to help the poor, say, even if one lacked any disposition to do so. I beg to disagree. I think one ought to have that (first-order) disposition, in order really to have that commitment, although, of course, one may fail to live up to it—that is, one may *on occasion* and *in clearly specifiable conditions*, fail to actualise it. There will be more on this in section 3.1.

[36] I don't think, contrary to what theorists often suppose, that acknowledging this will impair the characteristic authority of propositional attitudes' avowals. See fn. 76.

[37] Of course this might invite the idea—put to me by an anonymous reader—that at a later stage in their development they might have these very same mental states together with the ability to assess the evidence on which they are based and be able to offer it, if requested. Thus, commitments would just be a different mode of presentation of the very same kind of mental state originally held as a disposition. While I am comfortable with this idea in the case of some mental states as dispositions—those presently under scrutiny—, I don't find this view very plausible when the dispositions at stake are unconscious mental states of a Freudian kind. Whether this should be taken as a recommendation to split the category of mental dispositions at least in two, or else to resist the idea that commitments are ontologically identical to mental states as dispositions, even in the apparently harmless case, is not something I am able to expound on at present. Be that as it may, it seems to me that either way it is important to recognise the role of mental states as commitments, if only as different ways in which an ontologically unique kind of entity may be given to a subject, contrary to the dominant tendency in the philosophy of mind, whereby all (propositional) mental states are usually treated in a functionalist, hence purely dispositional way.

still explain why having beliefs and desires as commitments, that is to say, as brought about by judgment in the way described, should entail that they are known to a subject who has them.

Here's a first, partially unsuccessful shot:[38] in order to have beliefs and desires as commitments one should be able to withhold from them in case contrary evidence or countervailing considerations came up. So, for instance, to have the belief as a commitment that it is raining now one ought to withdraw from it if it were shown, say, that the street is wet because it has just been cleaned. Now, that information wouldn't require one to change one's mind if one were just imagining that it is raining now or were so hoping. So, that information can make one change one's mind just in case it is taken by a subject to bear on her *belief* that it is raining now. Hence, having mental states as commitments—that is to say, as mental states that are within one's control and for which one is rationally responsible—entails that one knows them because it is only if one does that one can actually have them. So, having mental states as commitments would entail that the subject who has them would know them.

This move, however, taken as such, is unsuccessful for (at least) two reasons. First, because I think we can at least conceive of a self-blind subject. Namely a subject who is capable of mental states as commitments, that is able to withdraw from her assertion of "It's raining now" in light of new counter-evidence, or that is able not to use that piece of information in her deliberations to action, if she got that new information—and yet, if asked "Do you believe that P?" would be unable to answer. Thus, while having beliefs as commitments, she would lack knowledge of them.[39] Secondly, because even if we had managed to establish the desired connection between commitments and self-knowledge, we would have merely shown that the latter is a *necessary condition* for the former, but we wouldn't have provided any *substantive* account of it.[40]

[38] This is Bilgrami's move (2006, pp. 160-166; 175-205). Notice, however, that I disagree with Bilgrami only insofar as he holds that the ability to engage in belief- (and other mental states-)revision *and nothing else* is sufficient for self-knowledge. As we shall see in §2.2, in my view, this is one necessary condition which, when combined with another one, adds up to a sufficient condition for self-knowledge.

[39] Self-blindness has been extensively discussed and criticised by Shoemaker 1996, pp. 226-245. Notice, however, that Shoemaker is taking for granted that a subject who *already has the relevant psychological concepts* can't be self-blind. As he writes very clearly in an earlier essay "[…] second-order belief, and the knowledge it typically embodies, is supervenient on first-order beliefs and desires—or, rather, it is supervenient on these plus a certain degree of rationality, intelligence *and conceptual mastery*" (Shoemaker 1988, p. 34. Emphasis added). Bilgrami thinks that self-blindness is inconceivable, when we are considering commitments, because of his specific understanding of that notion, as a second-order disposition to criticise oneself, or to accept to be criticised for not having lived up to them. Here, however, I have proposed a different account of commitments. But, at least on the face of it, it seems to me that one could criticise oneself or accept criticism from others for not having lived up to one's commitments without thereby having knowledge of them *as the mental states they are*. Consider a subject who has the desire as a commitment to help the poor, and is therefore able to judge "One should help the poor", give evidence for this, etc., but didn't have the concept of desire. Now, it seems to me perfectly conceivable that if someone told him "You are doing badly. You said the poor ought to be helped, but you aren't doing anything to that end", he would (have to) accept criticism, without thereby having knowledge of his desire *as such*. So I think self-blindness is metaphysically possible also for subjects capable of propositional attitudes as commitments. It is a further issue whether it is or may actually be instantiated.

[40] This criticism can in fact be leveled against Bilgrami's account of self-knowledge, although I suppose he would dig in his heals and insist that he has done all the explanatory work there is to be done, as long as he can redeem the left-to-right side of the constitutive thesis by placing it within the scope of an "agency" condition, complexly characterised as

2.2 Conceptual mastery as the result of blind drilling

In order to meet these objections, I think we have to introduce a further ingredient into the picture so that we will have two substantive[41] constraints on the C-conditions that, once *jointly fulfilled*, will make the left-to-right side of the biconditional hold as a matter of conceptual necessity. The missing ingredient—I submit—, beside the capacity of having propositional attitudes as commitments, is *conceptual mastery*. This may come as no surprise, since one would have imagined from the start that in order to judge and therefore believe that one *believes* (or desires/intends/wishes/hopes) that P, one would have had to possess the relevant psychological concepts. Hence, in my view, self-blindness is inconceivable only once we are dealing with a subject who is rational, i.e. capable of propositional attitudes as commitments, is sincere, cognitively lucid and, finally, equipped with the relevant psychological concepts.

However, I think it is only by giving a substantive account of what such a conceptual mastery may consist in that we can actually both avoid surreptitiously falling back into unwanted models of self-knowledge, and say something more meaningful about self-knowledge itself, in such a way that we won't be left with the impression that there is still some explanatory work to be done. So the question is: how does one conceptualise one's first-order mental states? Obviously, the answer can't be: either by having such states in view, as it were, and by recognising and labelling them as the mental states they are; or else, by *self-consciously* applying the rule that if I *judge* that P is the case (or is worth pursuing) on the basis of evidence, then I believe (or desire) that P. For, in the former case, we would be back with the observational model of self-knowledge, and, in the latter, we would presuppose knowledge of our own judgment, which is nothing but a mental state (or, in fact, an action). Furthermore, we would presuppose the possession of other intentional psychological concepts, such as the concept of judgment, which, arguably, will have to be explained along the same lines as the possession of the concept of belief (and desire). Finally, the self-conscious application of the introduction rule for the concept of belief (in this case) would presuppose the possession of the latter concept. Hence, the explanation would be hopelessly circular. It is therefore crucial to come up with a different account of what mastery of the concepts of belief and desire (in the first person present) consists in.[42]

---

involving reference to justifiable reactive attitudes (see Bilgrami 2006, p. 119), that reveal the connection between transparency and agency so characterised. Bar-On 2004, pp. 346-350 makes a similar critical point in connection with Wright's constitutive account.

[41] Recall that there are also two non-substantive constraints; namely: cognitive lucidity and sincerity (as we will see in section 3.3, sincerity is not entirely trivial).

[42] As to the possibility of having *tacit* knowledge of the conceptual role of the concept of belief, it must be noticed that it would still presuppose tacit knowledge of one's first-order judgment. How, then, would we account for this form of self-knowledge in its turn? I develop this objection in my 2008.

Here's a tentative view. Take a subject who is able to judge that P, give evidence in favour of it, use it as a premise in one's reasoning and withdraw from it if required, and has, therefore, the first-order belief as a commitment that P. Suppose you ask her "Do you *believe* that P?" and she is unable to answer. So you would conclude that she doesn't have the concept of belief. In which case, you could simply train her to the use of that verb by *drilling* her into using the *expression* "I believe that P".[43] Similarly, take a subject who says "My children's proper education is worth pursuing" and is disposed to offer considerations in its favour, does what she can to help bring that about, and withdraws from it if those considerations did (incredibly) no longer seem compelling; but if asked "Do you *desire* that your children should have proper education?" didn't know how to answer. Then again you could drill her to use "I desire that P" as an alternative expression of her mind—of her judging "My children's proper education is worth pursuing".

Let me stress that it is absolutely essential in order for the present proposal to steer away from any observational model of self-knowledge or from surreptitiously assuming such knowledge of our own mental states, that one should be adamant that "I believe/desire (intend/whish/hope) that P" are taught *blindly*: they are *ingrained* as an alternative way of expressing one's first-order beliefs and desires (and further propositional attitudes as commitments), other than by asserting "P", or "P is worth pursuing", or "It would be good (for me) that P be the case".[44] So, on the present account there would really be no inner epistemology—just a substitution of one form of behaviour with another. But—and this is crucial—the kind of behaviour which would get replaced would already be quite rich. For, in order to have beliefs (and other propositional attitudes) as commitments, a subject will already have to have the *ability* to differentiate between, for instance, believing P and P's being the case, by being sensitive to the fact that her point of view may be challenged—thus responding with reasons in favour of it—or proved wrong—thus abandoning it. It is only on the background of this already complex pattern of behaviour, which, however, doesn't seem to require the concept of belief (or of any other propositional attitude), but merely the capacity for first-order beliefs (and other propositional attitudes) *as commitments*, that I think we can maintain that "I believe that P" may be taught blindly. "I believe that P" would then be taught as an alternative way

[43] This account seems to me to be in keeping with Evans' point according to which all is needed to make self-ascriptions of belief is to judge that P is the case and preface that with "I believe that". Evans, however, isn't explicit about how the concept of belief in the first person present would be acquired. See his *The Varieties of Reference*, Oxford, Oxford University Press, 1982, pp. 225-226.

[44] Indeed, this seems to me to be the right development of Wittgenstein's idea that avowals substitute behaviour. It is just that when we move from avowals of sensations to avowals of propositional attitudes the behaviour we must take into account is not merely physical but also linguistic. I was pleased to find a similar point in Bar-On 2004, Ch. 7. However, Bar-On makes this point in the context of articulating a purely neo-expressivist account of psychological avowals that explicitly rejects the constitutive model. As we shall see in the following, I think this is just the beginning of an account of the role of "I believe/desire that P". (An expressivist account of Wittgenstein's views on avowals is given in Jacobsen, R. 1996 "Wittgenstein on self-knowledge and self-expression", *Philosophical Quarterly* 46, pp. 12-30, although, to my mind, it actually provides equal scope for a *constructivist* reading of Wittgenstein. See fn. 72).

of making the commitment to P other than judging that P. But what "I believe that P" would make *explicit*—to the subject herself and others—is the fact, which remains only implicit in judging P, and in forming the corresponding commitment, that that is just her own point of view among other possible ones, which need not be correct. This would happen by telling the subject, for instance, "See, you have said that P, but it is not the case that P. So you merely *believe* it".

An important feature of the present account is that it tights first-person and third-person uses of "to believe" together from the start. For it is only by being taught *by someone else* to replace the direct expression of one's mind—of one's beliefs as commitments—by means of asserting "P", while being disposed to retract it, if shown wrong, with appropriate psychological self-ascriptions, that one acquires the concept of belief, though not the ability of having beliefs as commitments. Once endowed with the capacity for making explicit her belief that P as a commitment, the subject can *then* articulate the conceptual role which individuates the concept of belief: for she can now express the difference between believing that P and its obtaining, both in her own case and in the third-person case. But the newly acquired ability to articulate that difference—which displays her conceptual mastery—shouldn't obscure the fact that she may well have been already practically sensitive to it and that her grasp of the concept of belief may not have depended on any substantial cognitive work.

2.2.1 Objections from empirical psychology

Coming from the psychology camp, various objections may be raised against this proposal; first and foremost that empirical evidence shows that children take time to acquire the concept of belief and that that goes hand in hand with the development of a theory of their own as well as of other minds.[45] This evidence wouldn't sit well with my proposal and would rather favour an account of concepts' possession, according to which to possess a concept—and, in particular, the concept of belief—consists in knowing its conceptual role.

In response it may be said that, quite apart from the conceptual problems that would pose, such as presupposing self-knowledge and the possession of a lot of intentional concepts, here I haven't tried to present a *psychological* theory of concepts' possession. After all, what I have suggested is simply how someone who is already able to have first-order mental states as *commitments* may come to acquire such a concept. It may be that young infants simply don't qualify. Nonetheless, I think it is an entirely empirical issue if the psychological data currently at our disposal, like the age

[45] The *locus classicus* is Gopnik, A. 1993 "How we know our minds: the illusion of first-person knowledge of intentionality", *Behavioural and Brain Sciences* 16, pp. 1-15, 90-101. Gopnik's paper gave rise to an enormous literature. However, it doesn't look as if the case of desires and other propositional attitudes has been studied as extensively as the one of beliefs.

at which children pass the false-belief test in their own case as well as in the case of other subjects,[46] should be taken to show that children take time to learn how to use "*I believe*" (and acquire the corresponding concept), or else should be taken to show that it takes time for them to become capable of *beliefs as commitments*. For, as I understand it, they come to pass the false-belief test in their own case when they actually understand that their own point of view about the world (or that of other subjects) may be wrong. This, I take it, is at least a necessary condition for having beliefs as commitments. Furthermore, the ability to pass the false-belief test in the case of others may be explained differently than by appeal to the fact that children would possess a theory of other minds. For it would be enough to explain their correct answers to suppose that they issue them as if they themselves were in the other person's shoes. So, the ability to pass the false-belief test need not show that children possess a theory of their own minds as well as of others'. In fact it may actually just be taken to prove that they are capable of first-order beliefs as commitments and to project themselves onto others and therefore issue the correct answer to the false-belief test, without thereby having any explicit knowledge of their own and other minds, which, on my view, crucially depends on the possession of the relevant psychological concepts.[47]

The reason why I think the data at our disposal don't tell us clearly what is the case is that—quite understandably—the experiments haven't been designed to test the possibility I am advocating here. For, usually, children will be exposed to talk in terms of belief when they are actually in the process of acquiring the ability to have beliefs as commitments. This may well have confused the issues: we may have mistaken the fact that it takes time for children to learn to have beliefs as commitments as a sign of the fact that it takes them time to acquire the concept of belief. Furthermore, we may have imputed that difficulty to the fact that mastery of that concept would depend on the acquisition of a theory of one's own as well as of others' minds. A more telling test, then, would be to look at children who haven't been exposed to psychological talk up to the age of 3 or 4 (which is the age at which they allegedly come to have a theory of the mind and the concept of belief); see if, around that age, they pass the false-belief test in their own case as well as in

[46] The test, first designed by Wimmer, H. and Perner, J. 1983 "Beliefs about beliefs. Representation and constraining function of wrong beliefs in young children's understanding of deception", *Cognition* 13, pp. 103-128, usually consists in showing children a Smarties' box. Asked what there is inside it, they answer "Smarties". They are then shown that there is a pencil instead and asked what another person, who hasn't seen the content of the box, would answer. They pass the test if they say "Smarties".

[47] I don't think this makes me automatically side with simulation theorists. For one thing, I am suggesting that children need not have the relevant psychological concepts, not even in the first person, in order to pass the false-belief test. Among simulation theorists, Robert Gordon's proposal avoids the attribution of psychological concepts to children. Simulative abilities, in his view, would be hard-wired and connected to the operations of so-called "mirror neurons". I myself am sceptical of the fact that mere appeal to hard-wired mechanisms can fully account for such cognitive abilities, though it can certainly manifest some of their material preconditions. I am also sceptical of his recent proposal (Gordon, R. 2007 "Ascent routines for propositional attitudes", *Synthese* 159, pp. 151-165) of explaining the acquisition of propositional attitudinal concepts through ascent routines that are based on the expressive, though dumb use of the very words that would later on come to signal the possession of those very concepts. I can't, however, pursue these points here.

others'—where, crucially, the test shouldn't be phrased in terms of beliefs—; and *then* introduce them to talk in terms of belief (and other propositional attitudes). If, at that stage, it actually takes them a short amount of time to learn how to use "I believe", then I think we would have shown that the account of concepts' possession I have been proposing would in fact be compatible with human psychological development.[48][49]

Furthermore, it has to be noted that the conceptual role of the concept of belief is what theorists of concepts offer as an *abstract* individuation of that concept that supervenes on a *practice* of its use which, however, may come about in different ways. In particular, since on my proposal the commitments undertaken by asserting "P", as an expression of one's belief as commitment, and by asserting "I believe that P" would actually be the same (save for the fact that the latter would make explicit what the former leaves implicit, namely that the assertion of "P" expresses one's own point of view that need not be correct), it may well be that the conceptual role of the concept of belief in the first person present specifies the rules for the use of that concept which, in practice, may have been acquired by becoming able to have first-order beliefs as commitments first and by then being blindly drilled to express them by prefacing one's assertion of "P" with "I believe that".

Assuming that what I have been proposing is along the right lines, it is perhaps worth-noticing that it is a consequence of the suggested account of conceptual mastery that while it may be an open issue if and to what an extent language is necessary in order to have first-order propositional attitudes as commitments, on my view language is indeed necessary for one's knowledge of them. So, while pre-linguistic (or non-linguistic) creatures might still have the former, I am committed to the view that only linguistic creatures can have self-knowledge.[50]

---

[48] Should this test prove impossible, one might see if there are languages which don't have talk in terms of belief and other propositional attitudes. If speakers of those languages were in fact capable of beliefs as commitments, as I think there is no reason to be sceptical of, we could then test how long it would take them to acquire the ability to express their minds by self-ascribing beliefs and other propositional attitudes in a different language which contained these devices.

[49] The apparently difficult case, for my own proposal, would in fact be constituted by autistic patients affected by Asperger syndrome. While they don't fail the false-belief test, they seem not to have a theory of their own minds, and to lack a theory of other people's minds. (Frith, U. and Happé, F. 1999 "Theory of mind and self-consciousness: what is it like to be autistic?", *Mind and Language* 14/1, pp. 1-22). But several things must be noted: 1) when we look at their reports, what they show is that subjects affected by this syndrome have different kinds of *experiences* particularly of speech, and different *sensations*, if not an altogether lack of painful sensations (pp. 15-18). None of this would show anything relevant with respect to their propositional attitudes and their knowledge of them. 2) The only report which has a bearing on this issue (from Donna Williams (1994)) in fact seems to imply that she didn't have, as an autistic child, desires as commitments (she writes (p. 15): "Autism had been there before I'd ever known a want of my own, so that my first 'wants' were copies of those seen in others (a lot of which came from TV)"). In such a case, it would not be surprising that they would have to gain knowledge of their own minds in a third-personal way and that this would require some kind of theory of other minds. Finally, all the data are based on personal reports and, obviously, this wouldn't have any bearing on the possibility of having, and of having knowledge of one's occurrent commitments. For much of what they say could actually be due to forms of self-interpretation.

[50] In fact I would be inclined to maintain that only linguistic creatures can have mental states as commitments, since that would require the ability to articulate and defend their basing reasons, at least in principle, while allowing that pre- or non-linguistic ones may have propositional attitudes as dispositions. I can't, however, pursue the point here.

In this connection, it is worth stressing that the psychological literature on self-knowledge in non-linguistic creatures I am aware of is both difficult to interpret and actually potentially irrelevant. For what has been tested particularly in chimpanzees is merely the ability to know others' *perceptions*, such as seeing, and not their propositional attitudes—let alone the highly specialised class of propositional attitudes as commitments I have been trying to make plausible so far. Moreover, these studies also show crucial discrepancies. For instance, researches conducted by Povinelli and his associates deny that chimpanzees have knowledge of other subjects' perceptions, while those conducted by Tomasello and his lab support the opposite interpretation.[51] So I think we can actually conclude that, at the present stage of the inquiry, the empirical data currently at our disposal have in fact no bearing on the issue of whether only linguistic creatures can have knowledge of their own propositional attitudes as *commitments*.

To recap and conclude: in order to account for transparency (and hence to exclude self-blindness), the C-conditions figuring in the constitutive thesis must include reference to a lucid and sincere subject, who is capable of having propositional attitudes as commitments and who is endowed with the relevant psychological concepts, acquired through blind drill.

3. Authority

Now our problem is: how can we account for free, as it were, for the claim that when a sincere and conceptually competent subject self-ascribes a mental state, she has it? And even before engaging in this task, what grounds would there be to accept that any sincere psychological self-ascription made by a conceptually endowed subject capable of the corresponding first-order mental states as commitments is correct? Aren't cases of self-deception, however rare they might be, just a clear counterexample to that half of the constitutive thesis? So, no matter how good our qualification of the C-conditions was in guarding against possible counter-examples to the left-to-right side of the constitutive thesis, its other direction doesn't seem to hold—let alone to hold as a matter of conceptual necessity. Thus, the constitutive thesis in its entirety would have to be rejected.

3.1 Self-deception

One might, with Wright,[52] add to the C-conditions that the subject shouldn't be self-deceived (or anyway, that it is reasonable to assume that she is not). But, quite apart from sounding an *ad hoc* move, it seems that the very possibility of self-deception would show that constitutive accounts

[51] See Povinelli, D. J. and Vonk, J. 2004 "We don't need a microscope to explore the chimpanzee's mind", *Mind and Language* 19/1, pp. 1-22; followed by an *Appendix* with replies to objections coming from the other camp, at pp. 24-28. Tomasello, M., Call, J. and Hare, B. 2003a "Chimpanzees understand psychological states—the question is which ones and to what an extent", *Trends in Cognitive Science* 7, pp. 153-156; 2003b "Chimpanzees versus humans: it's not that simple", *ivi*, pp. 239-240.

[52] Wright 1989a, pp. 200-201.

don't have much of a point: after all, how could one be mistaken about one's own immediately available mental states if not by somehow going wrong in identifying them? Wouldn't such room for error be compatible only with non-constitutive accounts of self-knowledge?[53] So, it would come as really good news if we could account for self-deception differently, thereby showing that its existence wouldn't constitute a threat to constitutive accounts.

Bilgrami[54] has come up with an idea that I find illuminating: on his view, self-deception is a case where a subject self-ascribes a mental state and has it as a *commitment*, yet she also has *another*, opposite unconscious mental state. The irrationality is brought about by the clash between her commitments and her dispositions (in Bilgrami's sense of this term).[55] So, for instance (the example is mine), take a jealous wife who openly and sincerely asserts with her friends that she believes that her husband is totally faithful to her—and has all the reasons in the world to do so—but, then, once at home, is often inquisitive, searches his belongings, etc. According to Bilgrami, what we should say is that she believes *as a commitment* that her husband is faithful—after all she is prepared to assert it with friends and has all the reasons to think so.[56] Still, she also has the unconscious belief, *as a disposition*, that he is unfaithful to her, which is operative in her inquisitive behaviour. So, she is "self-deceived" all right, in the sense that she sincerely avows a belief and partly behaves in ways that run contrary to it. Yet, it isn't the case that she has a false belief about her own beliefs. Rather she has two, different—both in nature and content—beliefs that give rise to her distinctively irrational behaviour.[57]

---

[53] Wright 2001b, p. 324 seems to me to underestimate the implications of allowing for cases of self-deception, and so I think does Heal 2002, p. 276.

[54] See Bilgrami 2006, pp. 140-157; 278-280. (Cf. also Stoneham, T. 1998 "On believing that I am thinking", *Proceedings of the Aristotelian Society* 98, pp. 125-144). It must be stressed, in order to avoid confusions, that I am endorsing Bilgrami's account of self-deception with respect to those mental states one would *not* attribute to oneself on the basis of inference and observation of one's own behaviour and further available mental states. I think in the latter cases one could make genuine mistakes and self-attribute mental states one doesn't really have. If, then, one were to restrict self-deception, properly so-conceived, only to these cases, as Wright suggested to me in conversation, then the fact that one might go astray in self-*interpreting* oneself would not represent a counterexample to the view that *non-observational* or *immediate* self-ascriptions of propositional attitudes aren't open to failures of authority. The authority of immediate attitudinal avowals would then remain unchallenged.

[55] Notice that, in contrast, the opposition between two incompatible commitments wouldn't count as a case of self-deception, but of overt, fully conscious conflict.

[56] Notice, however, that for Bilgrami this wouldn't be necessary. See fn. 34-35.

[57] One may object that there are also cases of "negative" self-deception. Cases, that is, in which a subject says "I don't believe that P" yet behaves in ways that are explainable only by attributing to her the belief that P. Stretching the example slightly, but just because that would help make the point more vividly, think of Pascal who would say "I don't believe that God exists (nor that he doesn't)" and yet would behave as an irreprehensible Christian. In this case one wouldn't be self-ascribing any belief. Hence, the only option seems to say that one falsely believes that one doesn't believe that God exists (nor that he doesn't). But I think we can recast the example in such a way that it ceases to be a counterexample to authority. Here's how. We could say that the avowal is still the expression of the subject's mental state. Namely, of her commitment to not using "God exists" (nor its negation) as a premise of her practical and theoretical reasoning, which runs against the disposition to behave as a kosher—if I may say so—Christian and thus use that belief as a premise of her practical reasoning. Bilgrami 2006, pp. 147-154 elaborates on this kind of difficulty in somewhat different terms, essentially because he is considering the matter from the point of view of a third party that attributes self-deception to a given subject. It seems to me that looking at a subject's own avowal *simpliciter* allows us to clarify what is going on in these apparently difficult cases in a simpler way.

Here are some considerations I can put forward in favour of Bilgrami's account of self-deception. First, I agree with Bilgrami that his account is better suited than its rival to explain cases of *motivated* self-deception, viz. cases in which self-deception is the outcome of a conflict in the subject between, say, believing that P and believing that not-P. In these cases, one of the two mental states gets suppressed, while the other is endorsed. Yet the former can remain operative in shaping (at least part of) the subject's behaviour and lead her to various forms of inconsistency. Secondly, and more generally, what makes us say that a subject is self-deceived is a *conflict* between her psychological self-ascriptions and (some other part of) her behaviour. Now, conflict is usually brought about by the fact that there are *two opposite parties* (or more) *at fight*, neither of which need be wrong, but be simply responding to different motivations and concerns. In the case at hand, it then makes sense to think that while one part of the subject's personality is entirely confident and mature, the other is full of insecurities, which lead her to be suspicious of the behaviour of those around her. Of course there may be reasons for both attitudes: on the one hand, the fully open and trust-worthy behaviour of the husband; and, on the other, a perhaps (well-)motivated sense of insecurity about one's own power to attract a person and to involve him in a stable relationship. Finally, suppose the subject realises, either through self-analysis, or through the aid of a therapist, that she has such an unconscious belief about her husband's infidelity. Now, if it were just a matter of realising her own mistake, she should simply correct her psychological self-ascriptions. After all, when I get to know that the wall I am looking at isn't red, but white and lit by a red light, I would immediately correct my belief—that is, I would *substitute* it with the new one. But clearly this is not what would happen in the case we are considering. For the subject would presumably still believe as a commitment that her husband is faithful to her. What she would (or, at any rate, should) do, rather, is to try and realign her behaviour with her commitments. Obviously, this can take a lot of time and personal effort and may indeed never fully succeed. For all these reasons, it seems to me that Bilgrami's account of self-deception is by far preferable to the traditional explanation of this phenomenon in terms of false psychological self-ascriptions. As a result, self-deception is entirely compatible with the fact that a subject is authoritative with respect to her own mental states, as long as it is clear that the mental states she is authoritative about are merely those as *commitments*.

3.2 Constructivism

Having dispensed with the counterexample to authority—indeed, with what is usually regarded as the only counterexample to it—let me turn to the problem of explaining why it holds. Recall that we want an account of authority that does not make it the result of any cognitive achievement. For any cognitive achievement may, in principle, go wrong and, therefore, there could be counterexamples

to authority. But we have just seen that there aren't any.[58] Indeed, there can't be any if we want to be serious about the fact that the biconditional holds as a matter of *conceptual necessity*. So, the account *must* dispense with the result-of-a-cognitive-achievement picture, *tout court*. For, so long as psychological self-ascriptions are seen as reports on one's own mental states, the question arises of whether they are true or false. Unless one is prepared to suppose that our cognitive faculties may be infallible,[59] one couldn't account for the claim that authority holds as a matter of conceptual necessity.

One—I think unsuccessful—attempt at explaining authority may consist in holding a purely expressivist position, whereby "I believe/desire/intend/wish/hope that P", being just expressions of one's propositional attitudes, would always be "correct". Better, by default they wouldn't be open to error, because they wouldn't be in the business of semantic evaluation.[60] However, this proposal still doesn't explain why *necessarily* if one asserts "I believe/desire that P" one is right about it. For, after all, one could be merely sounding off. Or else, if some kind of "seriousness" constraint were imposed, what would exclude, on this picture, that one may sometimes say "I believe/desire/intend/wish/hope that P" "spontaneously and in good faith"[61] and yet not have the corresponding propositional attitudes at all?[62]

Another strategy may consist in maintaining that since there aren't counterexamples to authority, any competent and sincere assertion of "I believe that P" (or of "I desire/intend/wish/hope that P") would entail that one has the corresponding first-order belief (or other propositional attitude, as a commitment). Still this would hardly be an explanation of *why* authority holds, but, rather, a simple acknowledgement, or a consequence, of the fact that it does.[63]

In fact I think the most promising way of explaining authority, in keeping, to some extent, with Wright's original proposal of conceiving of self-ascriptions of propositional attitudes as *judgement*

[58] This is not to say that one's own avowals of one's mental states as commitments are always correct but only that they are open to a very limited form of error: either they are incorrect because of conceptual incompetence, or because of slips of the tongue. (It remains an open issue, which I can't take up in this paper, whether these failures could have analogues in thought).

[59] An assumption that Descartes was happy to make but that wouldn't find many supporters nowadays.

[60] Notice that contemporary expressivists in this domain, like Bar-On (2004, Ch. 8) and Jacobsen 1996, do not deny that "I believe that P" has the meaning and the truth conditions it is usually taken to have, while not having the role of asserting that one has that belief but of expressing it. There will be more about this in the following section, although in the context of defending a performative account of avowals. That move, however, makes it even more difficult to explain why there should be a presumption that one's avowals be correct (cf. fn. 62).

[61] Heal 2002, p. 280.

[62] It is not by chance that expressivists are usually happy to account for self-deception along traditional lines. But, then, how can they account for authority? In particular, how can they hold that it is an a priori feature of our linguistic practice involving psychological avowals? Or, at any rate, that there is an "asymmetric presumption of truth governing them" (Bar-On 2004, p. 403)?—What would make such a presumption asymmetric with respect to the presumption of truth which we may grant to judgments that, for instance, are based on reliable observation or inference? If, in contrast, they maintain that authority is "only apparent—an illusion fostered by the descriptivist assumption that self-ascriptions express second-order beliefs about mental states" (Jacobsen 1996, p. 16), how can their position be reconciled with the fact that authority seems to be a constitutive feature of avowals?

[63] I think this would be Bilgrami's strategy. In this connection, see fn. 40.

dependent,[64] would consist in inverting the direction of fit and in holding the following—*constructivist*—picture: psychological self-ascriptions such as "I believe/desire/intend/wish/hope that P" do *bring into existence* the corresponding first-order mental states, e.g. the belief/desire/intention/wish/hope that P. On this model, there would be a sense in which it is literally true that we *make up* or *create* our minds. Moreover, since one's judgments would bring into existence the relevant first-order mental states, those judgments would necessarily be *true*. In fact, self-verifyingly so.[65] Furthermore, there would be no temptation to think that one should have the first-order mental state in view first, in order to make one's judgment or avowal, which would thus result in knowledge. For, if there is no mental state *before* making the relevant assertions or judgments, then of course there is nothing to know, or be aware of, in the first place, which should be tracked in judgment.

No doubt this proposal is going to meet objections. In what sense do we create mental states? How could asserting or judging "I believe/desire/intend/wish/hope that P" suffice to bring about the corresponding first-order mental state?—These are all legitimate worries, but I think this proposal has some considerable attractions too, once phrased a little differently. What should be claimed is *not* that we create *all* of our own mental states. For mental states as dispositions would be there independently of our ability to self-ascribe them. Still, there is a clear sense in which we do create our minds as we have been reviewing in the previous section. Namely, by *judging* that something is the case (or is worth pursuing or having) we do create our beliefs (and other propositional attitudes) *as commitments*. The crucial point is that when "I *believe/desire/intend/wish/hope* that P"—that is, the corresponding psychological self-ascriptions—are acquired along the lines I developed in the previous section, as ways of making the same commitments as the ones undertaken by judging and asserting "P" or "P is worth pursuing/having", while having in view reasons in favour of P (is worth pursuing/having), it then becomes possible to use "I believe/desire/intend/wish/hope that P" *en lieu* of asserting (or judging) that P (is worth pursuing or having), in order to form one's first-order belief or desire that P *directly*. The difference between forming the first-order mental state by means of the second-order judgment, instead of forming it by means of the first-order one, is just the fact that "I believe/desire/intend/wish/hope that P" makes explicit what the first-order judgement leaves implicit, namely that that is just one's own particular stand-point on P (or its being worth-pursuing or having). Hence, we can bring about the relevant first-order belief and other propositional attitudes (as commitments) *either* by judging that P is the case (or is worth-pursuing or having), *or* by judging "I believe/desire/intend/wish/hope that P" (while having in view reasons in favour of P

[64] See Wright 1989a.
[65] Also the corresponding assertions would have the same effect, as long as sincerity is granted. For an account of what this means, in the present context, see section 3.3.

(is worth pursuing/having). For, to repeat, given the role of the latter locutions (either in speech or in thought) and of "P (is worth pursuing or having)", I can commit myself to P (is worth pursuing or having), thus bringing about the corresponding first-order beliefs or other propositional attitudes, either by simply judging the latter; or else, by judging the former, thus simultaneously making explicit my commitment to P's being the case (or to its being worth pursuing/having).[66]

Before turning to a defence of the view that "I believe/desire/intend/wish/hope that P" are performatives (at least on occasion), let me address one possible objection to constructivism. Some theorists have argued against it on the grounds that it would entail the *unreality* of first-order mental states:[67] if such mental states do not pre-exist their self-ascription, then they don't have *real* and *independent* existence. Since this is implausible, constructivism is doomed from the start. In response, I think it should be stressed that my brand of constructivism entails only that mental states *as commitments* don't *necessarily* have independent existence of the corresponding second-order judgements. So, the kind of constructivism I am advocating allows for (first-order) mental states *as dispositions* to exist independently of the corresponding self-ascriptions. Moreover, it also allows for the conceivability of the independent existence of (first-order) beliefs and desires as *commitments*, when they are merely brought about by judging that P (is worth pursuing or having), by subjects who don't have the conceptual resources necessary to make the corresponding second-order judgement. Notice in fact that, as already remarked,[68] strong forms of constitutivism merely hold that self-ascriptions of propositional attitudes as commitments can (and often do) bring about the corresponding first-order mental states. They need not claim, however, that the latter can't possibly exist without the former, for instance in subjects who did not possess the relevant psychological concepts. Yet, strong (as well as weak) forms of constitutivism are united in claiming that, to suitably conceptually endowed subjects (who are also lucid and sincere), first-order propositional attitudes as commitments are transparently known. Hence, their occurrence is of a piece with subjects' awareness of them as the mental states they are with the content they have.

---

[66] Obviously it would always be available to one to justify one's judgment—"I believe/desire/intend/wish/hope that P", say—*ex post* by appealing to the fact that one's evidence allow(ed) one to judge that P (is worth pursuing or having), thus giving rise to one's belief/desire/intention/wish/hope that P (as a commitment). The possibility of giving such a justification for one's judgment "I believe/desire/intend/wish/hope that P", however, shouldn't obscure the fact that the commitment to P (is worth pursuing or having) was actually made by judging "I believe/desire/intend/wish/hope that P". My account of attitudinal avowals then explains why philosophers, most notably Wittgenstein in *The Philosophical Investigations* II, x, have been tempted to reduce "I believe that P" to "P". Their mistake was due to failing to see that the *contents* of those judgments are different, while their insight was to recognise that the *commitments* undertaken by making those judgments (or the corresponding assertions) are virtually the same. I take up the issue of Moore's paradox in Coliva, A. 2005 "Moore's paradox and commitments. On this very complicated concept of belief", in P. Leonardi and J-J. Açero (eds.) *Facets of Concepts*, Padova, Il Poligrafo, 2005, pp. 233-252.

[67] See, for instance, Bar-On 2004, pp. 412-413. Notice how Bar-On finds support for the untenability of this view in general, from its untenability in the case of those mental states "we share with non-human animals and pre-cognitive children". But judgment-dependency shouldn't, I think, be meant to apply to the latter cases. See also Heal 2002, p. 286.

[68] See fn. 18 and §2.2.1.

Be that as it may, it has to be stressed that the fact that the existence of certain beliefs and other propositional attitudes—those as commitments—is taken to depend on judging I believe/desire/intend/wish/hope that P does not make those mental states less *real*. The crucial point is that judgment-dependence is a claim about the *provenance* of first-order mental states, not about their (*un*)-*reality*. What I have been urging is that there are two kinds of judgments that can bring about the *same* result (i.e., a belief/desire/intention/wish/hope as a commitment), namely: either judgments that are outright about the world (or what is worth doing or having); or else, judgments that make explicit the particular stand-point from which the world is conceived to be thus-and-so.

It may sound surprising, if not altogether alarming that in my view first-order judgements and second-order ones play such an interlocking role. But to dissipate the resistance to such a view, consider that any judgment/assertion that P (is worth pursuing or having) is always made by *someone* and hence is necessarily the expression of a subject's point of view, even when its subject matter is, as it were, the world. By being drilled to the use of first-person present-tense psychological *vocabulary*, subjects are simply endowed with the means to make that "grammatical" fact *explicit*. Once they are conversant with that practice, they can use second-order judgements or assertions directly, as ways of forming the same commitments they would form by making the corresponding first-order ones. It then seems to me that the point of our psychological self-ascriptions is first and foremost to make explicit to ourselves and others the fact that the world, broadly conceived, is always described or assessed from a particular stand-point—one among potentially many. Their further performative role is simply a result of having being trained to take part in a linguistic practice where the same commitments can be undertaken in two different ways. This, I take it, is also the deep truth in Wittgenstein's and Wright's positions: psychological avowals—and their equivalents in thought—are the result of being trained to take part in a *linguistic practice*, and have their main point in making any participant aware of the fact that her specific point of view is just one among other possible ones.

3.3 Performatives and commitments

When understood in the way proposed, a judgment (or a sincere assertion[69]) such as "I believe/desire/intend/wish/hope that P" is like a *performative*, namely like "I promise to buy you an ice-cream", "I hereby thee wed", "I hereby name you so-and-so", etc: it makes a certain thing *happen*, for it does create the first-order propositional attitude as a commitment. Where this, to

[69] Notice the sincerity condition placed upon second-order assertions, which can't, however, be carried over to judgments, since judgments are—when made—necessarily sincere. This seems to me enough to dispel the worry, raised by an anonymous reader, that one may judge or assert "I believe that P" and yet not form the corresponding first-order mental state as a commitment. Of course what remains entirely possible is that I don't act on the basis of such a commitment and thus fail to actualise the connected disposition. But this is no objection to the view which is being proposed here.

repeat, is possible precisely because judging "I believe/desire/intend/wish/hope that P" becomes just an alternative way of undertaking the same commitments one would make by judging that P (is worth pursuing or having), save for the fact that the former kind of judgement would also make explicit what the latter leaves implicit, i.e. that the judgement just reflects a subject's own point of view.

Many have argued against such a view of psychological self-ascriptions by maintaining that it would commit one to the implausible claim that they would lack content and couldn't, therefore, be sensibly prefaced by negations, or be embedded in suppositions, and otherwise wider contexts.[70] But this objection, if sound at all,[71] could be raised only in the case of *implicit* performatives—that is, those which don't make explicit the kind of commitment one is undertaking. For *explicit* performatives, like "I *promise* to buy you an ice-cream" and "I *believe* that P", are speech-acts, which can have *more than one function at the time*: they can make things happen but they can also say what is being done by means of them.[72] In this latter sense, they would retain truth-evaluable content. For instance, "I promise to buy you an ice-cream" is both a way of *making the promise* of buying you an ice-cream and of *saying what I am doing*, i.e. promising to buy you an ice-cream. Of course, what I am saying could be *false*, since I could be insincere. Similarly, "I believe/desire/intend/wish/hope that P" would be both a way of forming the commitment that P (is worth pursuing/having) and of saying what I am doing. Moreover, what I am saying, i.e. that I believe/desire/intend/wish/hope that P (as a commitment), could be false, since I could in fact not be making that commitment and simply trying to fool you.

---

[70] The *locus classicus* is Geach, P. 1965 "Assertion", *The Philosophical Review* 74. Reprinted in J. F. Rosemberg and C. Travis (eds.) 1971 *Reading in the Philosophy of Language*, Englewood (NJ), Prentice-Hall, pp. 250-261.

[71] The counter is usually that they could retain *minimal assertoric content* as well as *minimal truth*. Accordingly, it would suffice for minimal assertoric content that performatives can be embedded in negations and suppositions, and that they can undergo the usual tense transformations. Obviously they can. For "I am not going to buy you an ice-cream", or "Suppose I bought you an ice-cream" and "I did buy you an ice-cream" are perfectly sensible things to say. Minimal assertoric content then pairs with *minimal truth*—with the idea that it is enough to qualify as a truth-predicate that some platitudes and, in particular negation and the T-schema, are respected. It would then be a further issue whether these statements should be taken as *descriptions*; or else, as *expressions of commitments one is (or isn't) undertaking thereby*. This strategy can be found in Hacker, P. 1986 *Insight and Illusion: Themes from the Philosophy of Wittgenstein*, Oxford, Clarendon Press, p. 90; as well as in Wright, C. 1992 *Truth and Objectivity*, p. 28; and in Jacobsen 1996. Hacker (1986, p. 298), however, denies that minimal assertoric content would be compatible with truth-evaluations. Be that as it may, when we consider judgments (and assertions) of "P" (is worth pursuing/having) we could distinguish between the *content* of the judgment (or of the assertion), which obviously is truth-assessable, and what is being done by means of the *act of judging it*. It is the act of judging that P that brings about the corresponding belief (or further propositional attitude) as a commitment, which can get *expressed* in one's assertion of "P (is worth pursuing/having)", although the *content* of one's assertion would remain P (is worth pursuing/having). A similar distinction can be found in Jacobsen 1996, p. 26 and in Bar-On 2004, pp. 251-264.

[72] A similar point can be found in Heal 2002, pp. 282-288. But also in Jacobsen 1996, pp. 23-28. Strangely enough, Jacobsen who, officially, sets out to characterise an expressivist account of avowals, in fact ends up defending the claim they are performatives. See Jacobsen 1996, pp. 26-28. So I find myself in agreement with much he says, although I would insist on the difference between expressivism and constructivism: on the former, first-order mental states are already there and get simply expressed by the relevant utterances; on the latter, in contrast, utterances (if sincere) bring about first-order mental states, at least in some cases.

However, whenever sincerity conditions are introduced, one might suspect that, perhaps surreptitiously, reference is being made to the fact that one's assertions or judgements should track one's pre-existing mental states. Hence, the whole point of the performative account I am proposing would be pre-empted, for its main contention is precisely that first-order mental states can be brought into existence by one's making the relevant psychological judgements or assertions, leaving no room for the idea of tracking a pre-existing mental reality. But, in effect, this account of what sincerity would consist in is not compelling. For the sincerity condition, in the case of performatives, just amounts to *one's lack of the intention to fool one's interlocutor*—and does not consist in a correspondence between one's pre-existing (first-order) mental states and utterances (or judgements about such mental states). By contrast, when I do wish to deceive my interlocutor it is not the case that I first check within myself whether I have the belief or the desire that P, find out I don't, say, and then say the opposite. Rather, I utter the performative sentence without respecting one of its felicity conditions—since I have *another* mental state, viz. the intention to fool you. Thus, although I utter a performative sentence, I don't thereby bring about the corresponding first-order mental state.[73] That is why "I believe/desire/intend/wish/hope that P" can be performatives and yet, on certain occasions be false, since what would make them true hasn't in fact been brought about. (Conversely, once the sincerity condition is satisfied, "I believe/desire/intend/wish/hope that P" is true because what would make it true has been brought about by that very judgement or assertion).

One frequent objection raised against performative accounts of avowals is that they would introduce a difference in *meaning* between the first-person, present-tense use of the relevant psychological verbs and their third-personal use (as well as their first-personal non-present-tensed use).[74] Since this seems absurd, one should reject such an account of psychological avowals. However, I don't think this objection is compelling. For, if the meaning of a word is what is offered as an *explanation of its meaning*,[75] then we will offer just one kind of explanation for "believe", "desire" and so on. For instance, that to believe that P means to be disposed, *ceteris paribus*, to use P as a premise in one's practical and theoretical reasoning; that in order to believe that P one needs evidence in favour of P, that if one believes that P, then, *ceteris paribus*, one will assert that P, etc. But nowhere did the performative account suggest that in the first person case things should be any different. After all, on the present proposal, all is being suggested is that the commitment to P can be *formed* by judging (or asserting) "I believe that P", where it is part and parcel of making such a

[73] Notice that here an asymmetry between assertions and judgments may arise. For the performative judgment can't be overridden. So, one way of putting the point is that when I am insincere I utter a performative sentence without making the corresponding judgment.
[74] See Geach 1965, p. 260.
[75] See also Jacobsen 1996, p. 18 for a similar point, which can be traced back to Wittgenstein (cf. *Philosophical Investigations*, §560).

commitment that one should have the kind of *dispositions* just mentioned.[76] So, allowing for the performative nature of first-personal, present-tense avowals merely entails that there are two ways in which a subject can come to have certain dispositions (or in fact ought to come to have them): either because she is somehow finds herself saddled with them; or else, as in the case at hand, because she actively brings them about—or tries to bring them about—as an implementation of her own deliberations. Obviously, on occasion, a subject could fail to behave accordingly, but one shouldn't confuse the fact that sometimes a disposition may not be actualised with the fact that there is no real or genuine disposition at all. To offer what I think is an instructive analogy: the fact that occasionally a crystal vase can be struck and not break doesn't show that it doesn't have the disposition to break if struck. So, the fact that sometimes one could fail to implement one's own deliberations doesn't show that one lacks the relevant disposition.

However, what this suggests, in its turn, is that there is another, perhaps more important distinction to be drawn—that is, a distinction between self-, as well as other-directed ascriptions of beliefs and other propositional attitudes as *commitments* and as *dispositions*. True, we can *bring about* commitments only for ourselves, by judging or asserting "I believe/desire/intend/wish/hope that P". Yet, we may nevertheless ascribe this kind of mental state to other people. Suppose you are listening to a subject who asserts that P, gives a lot of evidence in its favour, bets her own head on P, as it were, etc. When you then report on her by saying "S believes that P", obviously what you would be correctly attributing to her is a belief as a commitment. If, in contrast, you were interpreting S's behaviour by attributing to her a mental state she has never avowed (and might never be in a position to avow), which, however, would be helpful to *you* to make sense of what S is doing; or else you were engaging in deep analysis of a Freudian kind of her behaviour, then you would be attributing to her a belief as a *disposition*.

Furthermore, not all uses of "believe" (or other propositional attitudinal verbs) in the first person present are performatives. For, sometimes, the same judgment or assertion, can be used as a simple *description*, like when one finds out about one's own beliefs or desires through a process of self-interpretation. Conversely, both past and otherwise embedded uses of "believe" (or or other propositional attitudinal verbs) in the first person, although not themselves performatives—for they can't bring about a commitment—may nevertheless be an *ascription* of a *commitment* one had in

[76] Hence I wouldn't be too keen to endorse the kind of dilemma Wright thinks there is in the fact that, on the one hand, avowals are authoritative and, on the other, they self-ascribe a disposition whose obtaining is assessable from a third-personal point of view. See Wright, C. 1987 "On making up one's mind: Wittgenstein on intention", in P. Weingartner and G. Shurz (eds.) *Logic, Philosophy of Science and Epistemology: Proceedings of the XIth International Wittgenstein* Symposium, Vienna, Holder-Pickler-Tempsky. Reprinted in Wright 2001a, pp. 116-142, especially at pp. 122-123. Jacobsen 1996 and Heal 2002 too seem to be highly struck by this dilemma. The dilemma, however, seems to me very much a function of a simplistic description of the situation. So, it calls more for a dissolution—for an account of why it doesn't really stand—than for a positive solution.

the past, based on the memory of having made it, or, for instance, of a commitment one is supposing to be making.

So, what we witness here is a variety of uses of self- as well as other-directed psychological ascriptions: although it remains that we can bring about commitments only in our own case by means of first person present avowals, we can ascribe commitments both to ourselves and others and, moreover, we can ascribe both to ourselves and to others mental states as dispositions. The reason why it is so is simple: each of us can deliberate only for herself, but we can, and obviously do see also other people as deliberative agents—that is, we do know when they are making commitments as opposed to when they are simply acting on the basis of mental states they may be saddled with, but which are not the result of any deliberation of theirs. The same, however, applies to ourselves too, so we can report on previously made (and perhaps already abandoned) commitments or engage in the supposition of undertaking them; but we can also self-ascribe mental states as *dispositions*. What self-interpretation and psychoanalysis help us do is to acquire that kind of (third-personal) knowledge of the latter kind of mental states. What they can't do, however, is to turn us into deliberative agents with respect to them, for no amount of theoretical knowledge about ourselves will, by itself, ever transform the mental states we thereby become aware of into commitments.[77]

So what should be claimed—in keeping with one of the main theses of this paper—is that we have two different notions of belief and other propositional attitudes—those as commitments and those as dispositions—that cut across the first-person/third-person divide. Indeed we do explain their meanings differently, as we have seen in one of the previous sections of this paper (§2.1). Nevertheless, the propensity to see them as two different species of the same genus, instead of altogether different kinds of mental states—beliefs and "shbeliefs", say—and thus to talk, in both cases, of beliefs—specified "as commitments" or "as dispositions" respectively—would then be explainable by reference to the fact that both beliefs (and desires) as commitments and as dispositions could be responses to evidence (although only beliefs and desires as commitments would depend on actively assessing it) and could have similar effects at least on our *non-linguistic* behaviour: in the case of either kind of belief, a disposition to behave on the basis of P and, in either kind of desire or intention, a disposition to bring about P. What is relevantly different is the way in which the respective self-ascriptions are made. For when "I believe/desire/intend/wish/hope that P" is judged or sincerely asserted to make a commitment, it is a performative and brings about the corresponding first-order mental state. But the same words or mental content can also be a report on

[77] In fact they may make action and deliberation even more difficult to attain.

one's dispositions, known to oneself through observation and inference on one's own behaviour, hence in a third-personal way.[78]

A related point has to do with an oft-made observation that "the mark of the mental" would be the first-personal, present-tensed use of psychological verbs.[79] Well, on the face of it, this remark is simply wrong. For some psychological *self*-ascriptions are made in a *third*-personal way, as we have seen, and are such that one is self-ascribing a mental state as a disposition. By contrast, we have already noticed that there can be third-personal, present-tensed ascriptions of mental states which, while not performatives, are nevertheless ascriptions of commitments. What I think is distinctive about (most) adult human beings' mentality, then, is both that we can *make commitments* and that we can actually *see others as mental deliberative agents*—that is, as subjects who are capable of making commitments.

Finally, another objection often raised against performative accounts of avowals is that "P" and "I believe that P" would turn out to have the same *content*.[80] This objection, however, is wrong because, obviously, the truth-conditions of the two sentences are different: "It's raining (at *l* at *t*)" is true iff it's raining (at *l* at *t*); whereas "I believe it is raining (at *l* at *t*)" is true iff I believe it is raining (at *l* at *t*). Clearly these are quite independent states of affairs—it may be raining (at *l* at *t*) and I may be ignorant of it; or else, I may believe it is raining (at *l* at *t*) and be wrong. What, however, would be identical in the two cases, according to the performative account of (as we can now say) *some* uses of "I believe that P" (and of other avowals) are simply the *commitments* one would undertake by *judging* or *asserting* either.[81] This is why, at least certain occurrences of "I believe that P, but it isn't the case that P" would be *Moorean-paradoxical*. For when "I believe that P" is an expression of a commitment to P's truth, it would be (at least) irrational to commit oneself to P's falsity, *as well*.[82]

4. Conclusions

The constitutive account of our knowledge of our own propositional attitudes I have proposed, if correct at all, shows how constitutivism worth its name will have to take a rather radical

[78] Notice, moreover, that they can also be used to report a commitment previously undertaken. In such a case they are based on remembering having made such a commitment and certainly not on inspecting one's own mind, as it were. Also, it may be possible to find out about one's own dispositions in a third-personal way and self-ascribe them, and subsequently form a corresponding commitment. All this would require two different mental actions, though the final self-ascriptions may be identical in form.

[79] With the usual caveats having to do with difficulties in exegesis, Wittgenstein seems to have had this view, as Jacobsen 1996, pp. 14-17 reminds us of.

[80] This idea may be suggested by some of Geach's observations (1965, p. 259). Wittgenstein is obviously considered the chief-holder of this view, for his well-known remark that "I believe that P" is just a tentative assertion of "P". *Philosophical Investigations*, Oxford, Basil Blackwell, 1953, II, x, especially pp. 190-191.

[81] In order to undertake these commitments one obviously need not have the concept of a commitment. Nor should this claim be understood as implying that the content of "I believe that P", say, is "I commit myself to P".

[82] This is the account of Moore's paradox I gave in my 2005.

constructivist twist. This makes constitutivism a viable explanation of self-knowledge only for very specific and limited kinds of mental states we can enjoy—those as commitments—and, connectedly, only for specific kinds of self-ascriptions of propositional attitudes—those which amount, in fact, to performatives. These, in their turn, seem to me to be the only attitudinal self-ascriptions, which deserve to be called "avowals" and to be granted a distinctive authority, for they themselves do bring about those first-order mental states they are about. To my mind this is no sign of irrelevance, though. For what this long and winding road has led us to see is that we do have a variety not just of mental states, but also of propositional attitudes, as well as of ways of knowing them. Thus, for once, it is perhaps not mere rhetoric to conclude by saying that a full account of self-knowledge—that is, an account of the different ways in which we have knowledge of the varieties of mental states we can enjoy, as well as of the varieties of psychological self-ascriptions which will express that knowledge—will have to be deferred to another occasion.