Self-knowledge (But not: "Know Thyself")

Annalisa Coliva
April, 7 2004

Self-knowledge (But not: "Know Thyself")

If I ask you right now "What are you thinking?", or "What are you feeling?" you really don't seem to have a lot of work to do in order to be able to answer my question. Be sincere—nobody is going to hear you—and voice it out loud "What a bore, I have this paper to read for the Fellows' luncheon. It's in philosophy of mind—mmh, what is it?—; I can't wait to go back to my own stuff. Well, after all, life is almost perfect at the Italian Academy, I guess I can put up with this little chore". So easy: to know your own mind takes just one moment reflection and you can immediately pour out a whole series of thoughts, beliefs, desires, feelings, emotions and sensations.

Knowledge of your own current mental states is "easy" knowledge: knowledge that doesn't seem to require any cognitive work. Sure, self-knowledge in general is more complicated than that: at the very least the Socratic dictum will involve finding out what one is or isn't capable of doing. One will have to go through life, prove oneself and see what kind of person he or she is: what kind of character traits, virtues and vices one might have. That's why the Socratic injunction is of moral significance. And, inevitably, the kind of knowledge Socrates is talking about is knowledge that it takes time to achieve—in fact a life-time. In some important sense, one will know whom one is only at the very end of one's days, when decisions—important ones—and the pursuit of projects will be over. Up to then, there will always be room for surprise, for finding out something new about oneself, in fact, for re-inventing oneself, if one so wishes. But even confining our considerations to knowledge of our own mental states and not of our character traits, vices and virtues, things may be less humdrum than in our initial case. Freud has taught us that so many of our own mental states can be hidden from us: my competition with my mother for my father's attentions, in short, my Electra complex and my "desire to live a happy-face Oedipus"—as a psychoanalyst once very perceptively told me (without having ever read Schiffer,[1] I believe)—will be hidden from me.

Well, I'm afraid we'll have to leave all the interesting cases of self-knowledge aside. Here, I'll be concerned only with the dull kind of case: the immediate awareness each of us has of her current mental states. Even that will be more than enough for half-an-hour presentation. And just to alleviate your understandable disappointment, think it this way: to fulfill the moral obligation to know yourself or to find out about your own unconscious mental states, you need to be aware of your feelings, wishes, hopes and desires while performing the actions that will help you find out who you are, or that can lend themselves to psychoanalytic interpretation. So, for instance, if you feel constantly unhappy while fulfilling your marital obligations and, nonetheless, you keep going, as it were, you may well reveal a steady personality but also a lack of enthusiasm for your existential situation and, possibly, a lack of courage—that is, the courage needed to bring about a change in yourself or in the external conditions that may make you feel more identified with your own personal life. Or consider this case: it's my introspectively available joy at, say, seeing my mother's beauty fade away that could reveal—to me, as well as to my psychoanalyst—my hidden rivalry and animosity towards her, despite my

[1] For the non-philosophers: Stephen Schiffer (NYU) is always distinguishing between "happy-face"/"unhappy face" solutions to all sorts of paradoxes and philosophical troubles.

otherwise impeccable behavior towards her. So, "boring"-self-knowledge is boring all right, but is what we need to explain anyway if we want so much as be in position to take up the issue of the non-boring cases. So, please, bear with it.[2]

First of all, I'd like to bring to your attention a number of features which set "boring"-self-knowledge (hereafter, self-knowledge for short—and to avoid sounding too self-deprecating) aside from all other kinds of knowledge—not of self-knowledge, mind you, but of knowledge in general, like our knowledge of outer states of affairs, or of other people's mental states. These will be the constraints on any acceptable model of self-knowledge. So, keep them in mind. Then, I'd like to consider two naturalistically-oriented accounts of our knowledge of our own mental states. I will claim that neither of them works and that, ultimately, there is no scientifically acceptable account of self-knowledge. So, if anyone ever thought that science could explain everything (but I'm sure no scientist ever thought that, only naturalist philosophers did), bad news: science can't take care of this!

At this stage one might reasonably ask: ok, but why don't you just talk about your own views on self-knowledge? Well, partly because although I'm an analytic philosopher, I'm not totally oblivious of the *Zeitgeist*—as a non-analytic philosopher would say, *en lieu* of using the prosaic word "fashion". And the *Zeitgeist* comes down to this: if you really want to be serious about the mind nowadays, you have to look at what science—neuroscience, preferably—tells us about it. Anything else would be like obscurantism in the Dark Ages. Partly, because I don't have any anti-naturalist prejudice. If anything, I have a scientific background both regarding my education and my upbringing. One of my favorite past-times as a child was to make experiments in the kitchen with my mum—who, by the way, is a physicist—and, for a long time, I toyed with the idea of doing either physics or math at university. So if anti-naturalism has to be, it has to be because I've been rationally persuaded that there is no alternative.[3] Moreover, because, as a meta-pronouncement about philosophy (and notice the slight pragmatic

[2] Footnote for philosophers. As a matter of fact, here I'll be concerned mostly with knowledge of propositional attitudes, such as beliefs, desires, hopes and intentions, rather than with knowledge of non-attitudinal mental states such as sensations, feelings, emotions, perceptions, imaginings and memories. The reason behind this choice is that there are important differences among mental states which may well lead to different accounts of what knowledge of them amounts to. For instance, sensations, feelings and emotions don't have correctness conditions; can lack representational content and intentional objects. Perceptions and memories, in contrast, have correctness conditions, representational contents and intentional objects, but are passively enjoyed by a subject, contrary to beliefs, desires, wishes, hopes, etc. Imaginings, in contrast, don't have correctness conditions, and have both representational content and intentional objects. There will be more on the nature of attitudinal mental states in the following.
For an account of our knowledge of sensations and feelings that is in keeping with the one that I'll eventually offer for attitudinal mental states, see Wittgenstein, L. 1953 *Philosophical Investigations*, Oxford, Blackwell. Wright, C. 1989 "Wittgenstein's later philosophy of mind: sensation, privacy and intentions", *Journal of Philosophy* LXXVI, pp. 622-34, Bar-On, D. *Speaking My Mind. Expression and Self-Knowledge*, Oxford-New York, OUP, *forthcoming*. The case of perceptions and imaginings is tricky and deserves further analysis, while the case of memories can be straightforwardly incorporated in the account I'll give of attitudinal self-knowledge. Of course, for my account to be complete it will have to be extended in all these directions, but this falls outside the scope of this paper.

[3] By now, people will have realized that the real motivation behind my anti-naturalism is my long-standing Electra complex: given that my mum is a physicist … you'll fill in the details, since I'm sure you've heard this story millions of times.

contradiction in what follows), I think that philosophy is no ideology, but an activity of intellectual scrutiny that should produce rational conviction(!). So, part of the job is to discuss other philosophers' views and motivate one's own opinions against theirs.[4] Finally, because the kind of account of self-knowledge I favor is really a "last-resort": nobody would feel drawn to it at first-blush. It's only when you see that it's your only remaining option that—maybe reluctantly—you will endorse it. And just to give you a taste of it, the view of self-knowledge I favor—which, for the records, is a kind of constitutive view about self-knowledge—says something like this. Self-knowledge is no cognitive accomplishment at all. In fact, in an important sense, it is a misnomer to call our immediate awareness of our own mental states "knowledge". What goes by the name of "self-knowledge" is in fact a set of two conceptual truths that we can redeem by reflecting on the notion of rationality—of what it means to have certain kinds of mental states—and by granting subjects with conceptual capacities. So, the talk will have a substantial negative part because I can't, in fact nobody can do any better than this, to motivate the embracement of the account that I wish ultimately to maintain. So, here we go.

The constraints on any feasible account of self-knowledge

Our knowledge, in general, is either observational, or inferential, or a mixture of the two. Suppose I want to know what color the curtains of my lounge are. I take a look: pale yellow. Or suppose I want to know whether Socrates is mortal. I reason like this: Socrates is a man, all men are mortal, and therefore Socrates is mortal.[5] Finally, suppose I want to know where my father is, right now. I call up his secretary back in Italy. She tells me he's on a case now. So I'll infer that he is in court, since he is a lawyer.

Self-knowledge is just not like that. It isn't inferential because, sure, sometimes we infer our own mental states from other ones. Consider, for instance, this passage in Jane Austen's *Emma* in which our heroine realizes her love for Mr. Knightley—her long-lasting friend:

> Emma's eyes were instantly withdrawn; and she sat silently meditating in a fixed attitude, for a few minutes. A few minutes were sufficient for making her acquainted with her own heart. A mind like hers, once opening to suspicion, made rapid progress. She touched—she admitted—she acknowledged the whole truth. Why was it so much the worse that Harriet should be in love with Mr. Knightley than with Mr. Churchill? Why was the evil so dreadfully increased by Harriet's having some hope of return? It darted through her, with the speed of an arrow, that Mr. Knightley must marry no-one but herself.[6]

[4] As an aside: that's also partly why it is very hard to see how one could ever do philosophy without having knowledge of how (at least) some issues have been articulated throughout the history of this discipline. But I'm afraid my talk will have nothing historical to it.

[5] Of course there is also inductive and abductive reasoning. Still there is no need to expound on those, at this illustrative stage.

[6] Jane Austen, *Emma*, Penguin, London, 1987, 398, quoted in Wright, C. 1998, "Self-knowledge: the Wittgensteinian Legacy", in C. Wright, B. Smith, C. Macdonald 1998 *Knowing Our Own Minds*, Oxford, Clarendon Press, 15 borrowed from Tanney, J. 1996 "A constructivist picture of self-knowledge", *Philosophy*, 71, 405-22.

But, obviously, for the inference to get started at all, Emma must already have knowledge of those mental states of hers that figure as contents of the premises of her reasoning, such as, say, "I feel terrible about the fact that Mr. Knightley could return Harriet's feelings for him". On pain of an infinite regress,[7] there must be knowledge of our own mental states that isn't inferential. Somewhere down the line we must be able to know our own mental states directly.[8]

Quite intuitively, self-knowledge can't be a matter of observation either: first, mental states are just not the kind of things one could observe—as mental, they can't have any causal efficacy, so, they can't be perceived by the senses. Secondly, the Cartesian picture of an inner eye which is supposed to observe mental states that are luminously presented in the mental arena is more a recipe for trouble than a viable explanation of the sense in which self-knowledge could be observational. It's a recipe for trouble because it would lead to solipsism—the idea that each of us is caught up in her own world insulated from anyone else, being unable to know whether others have mental states and are, therefore, full-fledged persons. After all—the train of thought would be—how would I know that other people have mental states if those mental states are intrinsically private to them and foreclosed to me? It really becomes a recipe for catastrophe when taken a step further, because it would entail conceiving of our psychological language as private. But, since Wittgenstein,[9] private language has become synonymous of incoherence. The idea is this: any language is a rule-governed practice and it is essential to it that there be a distinction between correct and incorrect applications of the terms. But if the meanings of the psychological vocabulary are constituted by mental states that are private to each individual, then whatever seems to me the right application of the psychological term "S" is right—I am the master! And this just means that the distinction between being right/seeming right, which is much the same as the distinction between correct/incorrect applications of a word, has vanished into thin air. Yet, we have seen a moment ago that that distinction is essential to there being a (psychological) language at all. So, Cartesianism about self-knowledge takes us directly to incoherence: to a conception of our psychological language that turns it into a non-language.

"Wait a second"—you'll probably want to protest—we do have a psychological language: we tell each other about our own thoughts and feelings, sensations and emotions all the time and we seem to understand each others pretty well"—That's right! So, please, agree with me that Cartesianism won't do: throw it in your mental recycle bin. By the way, don't forget to throw with it also the idea that self-knowledge could be observational. Cartesianism was in fact the best way of cashing out the observational model. If Cartesianism has to go, so has the observational model. Hence, don't try to resurrect it. We gave it its best chance. It failed. Leave it in the bin.

Let's go back to my initial question: "What are you thinking?"—By now, if, indeed, you have managed to get to this point, you won't answer the same way. Nonetheless, you'll easily answer my question, with no effort and no work whatever.

[7] As a matter of fact most philosophers (one venerable exception seems to have been Nietzsche) hate the idea that one could go on and on in a circle forever.
[8] Cf. Wright, *op. cit.*, 16.
[9] Wittgenstein, L. 1953 *Philosophical Investigations*, Oxford, Blackwell.

Your mental states seem to be directly, or also, as philosophers find it desirable to say, transparently known to you. What they mean, I take it, is that their occurrence is of a piece with your awareness of them.

Surely, however, the occurrence of states of affairs out there—even the most banal ones around us—isn't of a piece with our awareness of them. The trivial fact that there is some pigeon flapping its wings right now in the middle of Amsterdam avenue isn't of a piece with my awareness of it. Nor is the occurrence of other people's mental states of a piece with our awareness of them. No matter how good I might be at figuring out what's crossing your mind right now, your feeling bored, being perplexed, or annoyed by what you're reading isn't something I'm immediately aware of. As I said, it's something I will have to figure out by taking into account your facial expressions, your sighs, your bodily movements, connect them with my general knowledge of what those reactions are an expression of and finally infer that, *ceteris paribus*, yes, you're bored, perplexed or annoyed. Transparency, then, is one of the features that set self-knowledge apart from all other kinds of knowledge.[10]

The other feature that sets self-knowledge apart from all others is this. If you are sincere and competent with respect to the concepts you use to express your mental states, nobody can—rationally—cast any doubt on your avowals. If you answer my question "What are you thinking?" by saying "I think this paper is getting more complicated than I thought it would be" and you are sincere and know pretty much how to use 'I', 'think', 'this', 'paper', …, then nobody could challenge you by saying "Are you sure that this is what you are thinking?", "How do you know it?", "Give me your grounds for your claim". As philosophers like to say, you are authoritative with respect to your own mental states: if you say (or believe) that you are thinking that this paper is getting more complicated than you thought it would be, then you are thinking that this paper is getting more complicated than you thought it would be. Punkt.

Surely, however, if someone had asked you "What's the weather like?" or "How is your mum?", then from your sincere and conceptually competent answers "It's raining" or "She is very sad since granny died recently" it wouldn't have followed at all that it would be inappropriate for someone to challenge you by saying, for instance, "Are you sure? You haven't taken a look out of the window and one of those relaxation-cds that make all the noises of nature is on". Or: "Are you sure? I've seen your mum last night at a party, dancing 'Staying alive' with your dad, she didn't look sad at all". So, authority is the second and final feature that sets self-knowledge apart from all other kinds of knowledge.

"Wait a moment"—you'll probably want to say—"it isn't true that any time you avow a given mental state then you have it. Take this case. There is a jealous wife: she sincerely says that she believes that her husband is faithful to her, but she also searches his belongings, is always inquisitive, and so on and so forth. So, surely, we are entitled to

[10] Probably the attentive reader may want to object that in the case of unconscious mental states transparency doesn't hold. That's right but, as noted at the beginning, we aren't trying to account for all kinds of self-knowledge, just for knowledge of occurrent, conscious mental states which will be needed anyway in order to gain inferential knowledge of our unconscious mental states. There will be more about this in the following.

cast doubt on her avowal and in fact conclude that, as a matter of fact, she believes that her husband is unfaithful to her".

Good, that's exactly right. You've put your finger on the so-called phenomenon of self-deception. Now, let me just notice at this stage that self-deception can't be the norm and that, therefore, it does not constitute a counterexample to the view that, at least most of the times, we do have authoritative knowledge of our own mental states. (This, of course, is not the end of the story, but it will do for the time being).[11]

To recap: In this section we have seen that self-knowledge is transparent, authoritative and can't be conceived as either inferential or observational. So how should we explain it?

A naturalist account of self-knowledge I: Armstrong

David Armstrong[12] is an Australian and Australians don't like metaphysical complications. So they tend to think that there is just one kind of stuff—physical stuff—and everything must be explained (or, at any rate, explainable) in causal-nomological terms. So here's what Armstrong thinks about self-knowledge.

Self-knowledge is the result of the operation of a reliable cognitive mechanism. That is to say, our brains are so wired that whenever there is a first-order mental state, such as my belief that there is a piece of paper in front of me right now, I am in a certain brain state; then the operation of a suitable physical mechanism brings about the occurrence of another physical state which corresponds to a second-order mental state, viz. the belief that I believe that there is a piece of paper in front of me right now. End of the story.

This model is marvelously simple. For one thing, it gets rid of the Cartesian idea that self-knowledge is a matter of inner observation: there simply isn't any observation going on here. There is just a hard-wired mechanism which, given a certain brain state, causally produces another one. For another, it accounts for transparency: the mechanism gets into operation whenever there is the relevant first-order mental state and produces the corresponding second-order belief. That's why we are immediately aware of our own mental states. Moreover, it explains authority because, after all, the mechanism is a reliable one. So the second-order beliefs that are produced by means of its proper operation are going to be right. Since they will be true and reliably produced, then they

[11] Footnote for philosophers. The rest of the story is this: massive self-deception would bring about an enormous mismatch between what one says and what one does. Such a colossal mismatch would cast into doubt the fact that the subject possesses the concepts that are necessary to avow (or judge that one has) the relevant mental states in the first place. For possession of the relevant psychological concepts is usually granted when no such mismatch occurs. For instance, a subject can be granted with the concept of pain just in case she avows pain in circumstances in which, *ceteris paribus*, she would also be disposed to show a corresponding pain behavior. Similarly, a subject can be granted with the possession of the concept of belief just in case she avows her belief that *p* in circumstances in which, *ceteris paribus*, she would also be disposed to use *p* as a premise of her practical and theoretical reasoning. But if self-deception were the norm, then a subject would avow pain or the belief that *p* when she wouldn't be disposed, *ceteris paribus*, to show any pain behavior, or to reason and act on the basis of *p*. Thus, generalized self-deception would cast doubt on the fact that the subject really possesses the conceptual capacities that are necessary in avowing (or in judging)—albeit erroneously—her own mental states in the first place. (There will be more on self-deception in the following).

[12] See Armstrong, D. 1968 *A Materialist Theory of the Mind*, London, Routledge.

will amount to knowledge.[13] Finally, this model is scientifically acceptable for it relies only on physical stuff—the brain—and on causal relations that, hopefully, will be subsumed under physical laws. It's just a matter of time: neuroscientists will find out where the mechanism is located and how exactly it works.

Well, as simple and plausible as it may sound, this model is far from being satisfactory. The first thing one may want to notice is that whenever a causal mechanism is involved, it may break down. So it would be a pure contingency that most of the times in which subjects say that they have a certain mental state, they do have it and that most of the times in which they have it, they know that they do. So, on this account, transparency and authority would be only contingent and a posteriori. But it seemed that they were necessary and a priori features of our knowledge of our own mental states, so, was it just an illusion? "That's right"—Armstrong would say—"It was just an illusion, boy".

Ok, but what about this? When a causal mechanism breaks down we do not think that people whose mechanism has broken down are irrational.[14] So, for example, if I suddenly become blind, I would often be wrong about what kind of things are in my surroundings, but I won't be deemed irrational—just plainly wrong. By contrast, suppose my self-knowledge mechanism breaks down: I'll avow mental states I don't have. So my actions, guided by my first-order mental states, will obviously contrast with my avowals. For instance, I'll say "I hope mum will recover" and then systematically fail to buy the medicines the doctor prescribed for her. I'll say "I hope she gets better" and then freeze the room where she sleeps, and so on and so forth. What people would say—since I'm not lying—is that I'm irrational: I say certain things and do the opposite.

Well, maybe, but we are all acquainted with retarded people or with senile ones. They say certain things and do the opposite. We just pay no attention to whatever they say because they are not reliable indicators. Sure, their being massively wrong casts doubt onto their rationality but the basis for that verdict is the fact that we take them to be wrong. And as long as their being wrong can be explained by saying that their self-knowledge mechanism doesn't work no doubt has been cast on the adequacy of Armstrong's model.

Ok, right. Let's try this, then. Armstrong's model should be, at least in principle, empirically testable, since it is supposed to be scientifically amenable. So, suppose you want to find out the empirical correlation between a given first order mental state and the belief or claim that one has it. First of all, you must find out what neural states correspond to the first order mental states. This, however, can be done on the basis of purely behavioral criteria only in a limited number of cases. It could be done for pain, for instance—by assuming that if someone is physically injured and screams and moans then she is in pain. So, when you see those symptoms, you look into the subject's brain—with appropriate instruments, of course—and find out what neural state she is in. But, obviously, you can't do this for, say, the hope that peace should be reached in the Middle East: there are no merely behavioral criteria that correspond to hoping that peace should

[13] —on this externalist, anti-Platonistic account of knowledge according to which knowledge isn't justified true belief, but, merely, true, reliably formed belief.
[14] This is a suggestion made in conversation by Akeel Bilgrami.

be reached in the Middle East.[15] So, while in the case of pain one may individuate the neural configuration that realizes pain, independently of the avowal of pain and, then, verify whether a subsequent avowal of pain is reliably caused by the neural realization of pain, in the case of the desire that peace should be reached in the Middle East this cannot be done because, to repeat, we can't have access to the first order mental state independently of the subject's avowal that she has it. Hence, Armstrong's model, which promised to be scientifically amenable, fails to be empirically testable and fails to be so because it must presuppose self-knowledge (and indeed in the form of an avowal) in order to individuate the relevant first-order mental states in the first place. Too bad![16]

A naturalist account of self-knowledge II: Gopnik's theory-theory

Let's now turn to a genuinely scientific model of self-knowledge that has been developed by the American psychologist Alison Gopnik.[17] In some ways, it is a development of the old-fashioned and, by now, fallen into disrepute behaviorist approach insofar as it claims that self-knowledge is wholly inferential and the basis for the inference is (mainly) overt behavior. Yet, in some ways, it also incorporates the Cartesian point that self-knowledge is observational. The idea is that children around the age of three/four come to possess a theory of the mind on the basis of which they interpret their own behavior as well as others'. The application of this theory gives them knowledge of their own minds as well as of others'. So, basically, the idea is that they acquire knowledge of their own mental states, such as, for instance, one's desire to have an ice-cream, reasoning pretty much as follows: "Since I'm feeling hungry and I'm going towards the fridge where I believe there's an ice-cream, I desire an ice-cream".

Of course Gopnik is well aware of the fact that it doesn't seem to us to be doing any inference when we get to know our own mental states, even less that we have to wait and see how we behave in order to find out what we think. But she thinks she can account for the distinctive phenomenology of self-knowledge. Her idea is that like a seasoned scientist can actually see an electron in a cloud-chamber, similarly those who have acquired the theory (of the mind) will be able to apply it so rapidly that it will naturally seem to them as if they were actually seeing their own mental states. Transparency is therefore an illusion, or, better, a by-product of the practice of applying the theory to oneself. Authority, in contrast is just a straightforward illusion: there is no reason why, in principle, we couldn't be wrong about our own mental states, pretty much as we can be

[15] Or, at any rate, behavioral manifestations underdetermine the individuation of precisely that mental state.
[16] So, although there may be causal mechanisms that enable self-knowledge, as I presume there must be causal mechanisms that enable thought in the first place, self-knowledge can't be explained by appealing to them (nor thought, in my view. But this is another story). Notice the relevance of the demise of Armostrong's model: since it was the only possible development of the observational model, its failure entails that one of the two possibilities that seemed initially open to defend the view that self-knowledge is indeed knowledge is foreclosed for good.
[17] Gopnik, A. 1983 "How we know our minds: The illusion of first-person knowledge of intentionality", *Brain and Behavioral Sciences* 16, 1-14. Reprinted in Goldman, A. 1993 *Readings in Philosophy and Cognitive Science*, Cambridge (Mass.), MIT.

wrong about other people's mental states.[18] Still, it is true that with practice and being constantly around ourselves we become very good at figuring out our own mental states. That's why we are mostly right about them.

Well, there are various things to say about this model. First, this whole idea of seeing our own mental states is a muddle. Indeed, Descartes got the phenomenology of self-knowledge wrong: we don't see our own mental states. The use of that verb is highly metaphorical: we are just immediately aware of them. So the analogy between the child and the seasoned scientist is wholly beside the point—simply, there is nothing like seeing which should be accounted for in the first place. To repeat, when philosophers talk about transparency they are not—or at least, not necessarily—talking about mental states as objects that are directly visible. Rather, they are metaphorically referring to the fact that their occurrence is of a piece with our awareness of them.[19]

Secondly, and more importantly, even in the little sketch of reasoning I presented as an exemplification of what Gopnik has in mind, there are already specimens of self-knowledge, viz. the subject's knowledge of her feeling hungry and of her believing that there is an ice cream in the fridge.[20] So, knowledge of our own mental states seems to be presupposed rather than explained by the theory in order to have the necessary premises that should inferentially lead to our knowledge of our own mental states. To put the point in more general terms: you can't figure out a mental state starting from merely behavioral symptoms. So, for instance, you can't figure out your own desire to eat an ice cream starting just from your moving towards the fridge. You could be moving towards the fridge because it's you favorite place in the kitchen, because the dog is barking at you claiming his supper, and so on and so forth. So, in addition, you need to have access to your own beliefs, desires and other mental states in general which would explain your moving towards the fridge as an action directed to the goal of fulfilling your desire to have an ice cream. So, you can infer that you desire an ice cream because you are aware of your feeling hungry and of your moving towards the fridge because you believe that there is an ice cream there and that the ice cream will satisfy your appetite.

Thus, to conclude: the theory-theory is not an adequate model of self-knowledge because either it falls into a crude and flawed form of behaviorism; or else, it presupposes self-knowledge.[21]

---

[18] And indeed the experimental evidence on which Gopnik develops her model is precisely pointing to the fact that before a certain age children just make a lot of mistakes both in the ascription of mental states to others and to themselves.

[19] So I wholly agree with Richard Moran (*Authority and Estrangement*, Princeton, Princeton Univeristy Press, 2001, 14) who writes: "While 'representationalism' is a controversial thesis about the ordinary perception of objects in the world, on nobody's view is the awareness of one's headache mediated by an appearance *of* the headache. And in the case of attitudes like belief, there is simply nothing quasi-experiential in the offing to begin with. There is nothing it is like to have the belief that Wagner died happy or to be introspectively aware that this is one's belief, and that difference does not sit well with the perceptual analogy".

[20] To be fair, Gopnik allows for non-intentional mental states to enter the inference. But the point is that you need also intentional ones, such as your belief that there is an ice cream in fridge, to perform the relevant inference.

[21] Footnote for cognitive scientists. Simulation theorists such as Alvin Goldman have claimed that the theory-theory has also the pressing problem of explaining where the psychological concepts children apply to themselves and others come from. The thought is that even if they are often wrong, up to a certain age, in

A constitutive account of self-knowledge

The constitutive thesis

So far we have seen that self-knowledge is neither based on observation, nor on inference and since we don't have any other way of knowing truths, we should conclude—with Crispin Wright and Paul Boghossian[22]—that it is based on nothing. What this means is that so-called self-knowledge is not a kind of cognitive achievement after all and it is somehow a misnomer to call it "knowledge" if knowledge is understood as the result of a however minimal cognitive endeavor. Rather, what we call "self-knowledge"—that is the distinctive kind of authority we recognize 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.

To illustrate this kind of position in more detail, all constitutive theorists agree that the following thesis is a priori true.

> Constitutive Thesis: if one has a first-order mental state M with content that *p*, then one will also be in a position to judge that one does, and if one judges that one has the mental state M with content that *p*, then one has it.[23]

What constitutivists debate among them is: what the grounds of the constitutive thesis are—e.g. is it grounded on the notion of rationality? Or, rather, on that of deliberative agency, etc.?—; and how to interpret the thesis and, in particular, what kind of metaphysical implications it has.

Here I won't have time to look at the details of various constitutive accounts and I will just present my own tentative and still underdeveloped proposal. As will become apparent, I suggest that the ground for the a priori truth of the thesis lies in our conception of rationality—of what it means to have certain kinds of mental states—and in granting

---

their applications of those concepts to themselves (and to others), they do have them. Where do these concepts come from, then? They suggest that they should come from one's own immediate awareness of one's own mental states and get projected onto others by means of simulation. Roughly, the idea is this: if I see someone moan and cry while injured I will simulate being in the same kind of state and apply the concept of pain by analogy to what I know I would be feeling if I was in that state. Contemporary simulation theorists corroborate their theoretical position by appealing to recent findings by two Italian neuroscientists, that is Rizzolatti and Gallese, according to which there are so-called "mirror-neurons" which fire when one sees someone else perform a given action and would also fire when one is performing that action oneself. Yet in the former case, suitable neural mechanisms inhibit the action.
A discussion of simulation theories falls outside the scope of this paper. Let me just remark, in passing, that, first, they are a contemporary version of the argument from analogy whose pitfalls have been shown—conclusively, I believe—by Wittgenstein, in those sections of the *Philosophical Investigations* concerned with knowledge of other minds. Secondly, that simulation theorists explain knowledge of our own minds pretty much along Cartesian lines and will, therefore, inevitably fall pray of Wittgenstein's attack on the very possibility of a private language.

[22] See Wright, C. 1989 "Wittgenstein's later philosophy of mind: sensations, privacy and intention", *Journal of Philosophy* LXXVI, pp. 622-34 (p. 631 in particular) and Boghossian, P. 1989 "Content and self-knowledge", *Philosophical Topics* 17, pp. 5-26 (p. 5 in particular).

[23] Often, the thesis is presented in the form of a biconditional: S believes/desires that *p* iff S believes that she believes/desires that *p*. Nothing relevant hinges on the style of the formulation.

subjects with conceptual capacities. My views on the metaphysical import of the thesis, in contrast, will be confined to the first appendix, as they involve some complicated purely philosophical discussion. (Of course, feel free to browse if you so wish).

The first half of the constitutive thesis: transparency

Let's focus on the first part of the thesis, which basically elevates the transparency of mental states to the rank of a conceptual truth and which is bound to generate some reactions. Two are most likely.

"Look,"—one may want to say—"what about unconscious mental states? If you allow for them, then they would be there even if one is in no position to self-ascribe them". That's right. So I owe you an answer. "Furthermore"—one might like to add—"we ascribe mental states to animals and infants to explain their purposive behavior that can't be explained simply in a causal-nomological manner. Still, we don't want to say that they have knowledge of their own mental states". Again, I think this is a fair point. So, I owe you two answers.

Let me start with the second objection. One strategy is to say that animals and infants don't really have beliefs and desires, just proto-beliefs and proto-desires. Full-fledged intentional mental states are the ones we have; they just have second-rate ones.[24]

Well, maybe so but think it this way: you see your dog go to the fridge at lunchtime and look anxiously at it. You then see your father do the same (—and, being an Italian father, he won't open the fridge, of course. Mum will have to do it for him). Well, I guess we would feel inclined to say that they both desire to have lunch and believe that there is food in the fridge. So, on the face of it, no matter how much simpler and less comprehensive a dog's overall belief-and-desire-system is going to be,[25] it seems to be composed of beliefs and desires all right (just like my father's).

Another—more promising—strategy is to point out that our notion of an intentional mental state isn't univocal. On the one hand, there are intentional mental states as dispositions: states that are attributed to the subject to make sense of her observable behavior,[26] which she may not be aware of and, even if she were, wouldn't be within her direct control.[27] If animals have mental states, then they have them *qua* dispositions. Moreover, unconscious mental states would also fall into this category. So, we can grant that the first half of the constitutive thesis does not hold for mental states as dispositions.

---

[24] Bilgrami in some moods seems to maintain this. Michael Dummett has repeatedly maintained the same view but on the grounds that animals, lacking full-fledged concepts, can have only proto-thoughts. Therefore, their intentional mental states having such proto-contents could only be proto-attitudes.

[25] My father, besides wanting lunch, will also have a definite conception of what he feels like having, how it should be cooked, etc. The dog, probably, won't.

[26] Of course, sometimes the interpreter may be identical to the interpretee, as Emma's example vividly showed. In that case the subject would become aware of her own mental states by reflecting on her behavior but wouldn't be directly aware of them. Said otherwise, she would gain knowledge of her own mental states in ways directly parallel to those in which one knows about other people's mental states.

[27] Knowledge of these states expressible in phrases such as "I'll explode if he goes on shouting at me" would be gained in a third-person way. That is to say, it is a prediction about oneself, based on knowing that one is a person who generally reacts against those who shout at her.

Yet, manifestly, adult human beings have also a further kind of mental states, namely, mental states that are within their control and for which they are rationally responsible. Call them "mental states as rational commitments".[28] Mental states as rational commitments are the result of judging that something is the case. So, for instance, my belief as a commitment that it is raining now is the result of having judged that it is raining now.[29] Similarly, my desire as a commitment that peace should be reached in the Middle East is the result of having judged that peace in that area would be beneficial. Characteristically, judgments—which are mental actions—are made on the basis of evidence and/or practical considerations. So, for instance, my judgment that it is raining now will be grounded on evidence, like looking out of the window and seeing the rain fall down and people walk by with their umbrellas open. Similarly, my judgment that peace in the Middle East would be beneficial will be based on believing that stopping the hostilities in that area will bring stability to it and will remove one of the causes of terrorism and that these are generally good things.[30] Since beliefs and desires as commitments are brought about by one's judgment, they are within the subject's own control: she could have not judged these things to be the case, but she did. Moreover, since the relevant judgments are made on the basis of evidence and of practical considerations, the ensuing beliefs and desires will be something for which the subject is rationally responsible. So, for instance, were it to be shown that the evidence invoked in favor of (the judgment that leads to) one's belief that it is raining now is defeated, or that the practical considerations invoked in favor of (the judgment that leads to) one's desire (as a commitment) that peace should be reached in the Middle East are overturned by further considerations, a rational subject ought to[31] withhold from believing that it is raining, or from desiring that peace should be reached in the Middle East.

Now, let's go back to the apparent counterexamples and consider the following. Animals don't have mental states as a result of judgment and of bringing evidence and practical considerations to bear on what they think. Hence, they can't have mental states as commitments. Similarly, unconscious mental states aren't, obviously, brought about by judgment. Hence, they aren't commitments. Thus, the cases of animals' and of unconscious mental states won't be counterexamples to the first half of the constitutive thesis. For that thesis holds only for mental states as commitments. Yet, we must still explain why having beliefs and desires as commitments, that is to say, as brought about

[28] Besides helping to account for self-knowledge, the view that we also have mental states as commitments helps explain Moore's Paradox, namely the paradox of saying (or judging) "I believe that *p*, but it isn't the case that *p*". See my "Moore's Paradox and commitments—Or on this complicated concept of belief", forthcoming in P. Leonardi (ed.) 2004 *Concepts*, Padova, Il Poligrafo.

[29] It is a conceptual truth that if I judge that p (or that p is good to have) then I form the belief (or the desire) that *p* as a commitment.

[30] Let me stress that judgments are fundamental and much more widespread than we might think at first glance. For, although the phenomenology of thought is such that it doesn't seem to us as if we are continuously engaging in judgments, the fact remains that any time we take evidence on board, like when we see the rain fall out of the window and form the corresponding occurrent belief, there is a maybe tacit, or implicit act of acceptance of it.

[31] That's why, for instance, wishful thinking is taken to be a form of irrationality: even if it turned out to be true, a father's stubborn belief that his son is alive, despite the fact that he disappeared in Vietnam, is held on the basis of no evidence, or of extremely weak one. Hence, it ought not to be believed.

by judgment in the way described, should entail that they are known to the subject who has them.

Here's a first—unsuccessful—shot: 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 in case it were shown that a film is being shot outside one's windows and the rain is a fake and the people walking by are cleverly disguised actors. Now, that information wouldn't require you to change your mind if you were just imagining that it is raining now or were hoping that it is raining now. So, that information can make you change your mind just in case it is taken by you to bear on your belief that it is raining now. Hence, having mental states as commitments—that is to say, as mental states that are within your control and for which you are rationally responsible—entails that you know them because it is only if you do that you can actually have them. So, having mental states as commitments entails that the subject who has them should know them. Since being capable of mental states as commitments is essentially equivalent to being capable of rational thought, we can put the same point by saying that rationality entails self-knowledge.

Still, some philosophers think that rationality *per se* is not sufficient for self-knowledge.[32] The idea is that we could just have certain mental states such as beliefs, even as a result of judgments, and withhold from them, if counterevidence came up (or even back them up with reasons), without having knowledge of our own mental states. Basically, what they are envisaging is a subject who is able to judge "It's raining", and who, on request, is able to say why, and who is such as to no longer judge that it is raining if her evidence is defeated and yet, if asked "Do you believe that it is raining?", is unable to answer, because—so the story goes—she would not be monitoring her own mental states. So, rationality, or, equivalently, mental states as commitments would be there all right, yet self-knowledge could be missing, contrary to the claim that rationality suffices for self-knowledge.

Furthermore—these philosophers protest—even if self-knowledge were a fall-out of rationality—understood in the way suggested—it would be just a necessary condition for having mental states as commitments, so an independent account of it should still be provided.[33]

I think both these objections are perfectly sound. So, in order to meet the first, we have to introduce a further ingredient into the picture so that we will have two conditions that, once jointly fulfilled, will suffice for self-knowledge. The missing ingredient—I submit—is conceptual mastery. Moreover, I hold that once we characterize properly what

[32] See, for instance, Moran, *op. cit*., pp. 107-113; Bilgrami, *A. Self-Knowledge and Resentment*. *How to Reduce Four Mysteries to One*, Boston, Harvard University Press, forthcoming, Ch. 4, *infra* (Bilgrami considers this position at length, although in the context of rejecting it); and Wright, in conversation.

[33] For a similar objection, see Bar-On, *op. cit*., Ch. 9 *infra*. Bilgrami is well aware of this objection and insists that he is not trying to give a reductive explanation of self-knowledge but only to place it within a suitable network of concepts that should help to clarify it. In fact, he is offering a transcendental argument for the existence of self-knowledge that goes like this: since self-knowledge is a necessary condition for responsible agency and we are responsible agents, then we have self-knowledge. My philosophical sensibility here is different: it seems to me that even if no reductive explanation is given, some more substantive account of what knowledge of one's own mental states consists in is required.

it takes to master the concept of belief (or desire) in the first person present, we will have also accounted for self-knowledge in a more substantive way, which won't leave us with the impression that there is still some explanatory work to be done. Let me expound on this a bit.

In order to believe that one believes (desires) that *p* nothing more is needed than being able to judge that *p* (or that *p* is good to have), being rational and conceptually competent.[34] For, suppose, that I judge *p* to be the case, I am rational and master the use of the concept of belief in the first person present. Then it seems inconceivable that I could fail to judge—if prompted—that I believe that *p*. So far so good, but the problem is: how do I conceptualize my first-order mental state? And the answer can't be: either by having the first-order mental state in view, as it were; or else, by applying the rule that if I judge that *p* is the case on the basis of evidence, then I believe that *p*. For, in the former case, we are back with the observational model, and, in the latter, we do presuppose knowledge of our own judgment, which is nothing but a mental state/action. So, we wouldn't have advanced in a bit.

Hence, it is crucial to come up with a different account of what mastery of the concept of belief (in the first person present) consists in. Here's a tentative view. Take a subject who is able to judge that *p* and give evidence in favor of it, 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 will conclude that she doesn't have the concept of belief (as the concept of an occurrent state of mind of hers). In which case, you simply train her to the use of that verb by drilling her into using the expression "I believe that *p*". And when I say that you should drill her into the use of that expression, I mean it: you teach her to substitute one form of behavior—one kind of expression of her mind, viz. the outright assertion of "p" accompanied by the ability to give reasons for it, which manifests her first-order belief—with another, viz. the assertion of "I believe that *p*".

Take then a subject who says, "Peace in the Middle East would be good to have" and is disposed to offer considerations in its favor, but if asked "Do you hope/desire that peace should be reached in the Middle East?" didn't know how to answer. Then again you drill her to use "I hope/desire that *p*" as an alternative expression of her mind, viz. of her asserting that "Peace in the Middle East would be good to have" for this and that reason.

Let me stress that it is absolutely essential in order for my proposal to steer away from any observational/inferential model that one should be adamant that "I believe that *p*", or "I desire that *q*" are not taught on the basis of evidence, which would presuppose, once more, knowledge of one's own mental states, nor on the basis of the rule "if you are disposed to do thus-and-so, then you believe/desire that *p*", which, again, would require knowledge of one's mental states, or, alternatively, the observation of one's overt behavior—all models which we have discarded in the previous sections. Rather, I must insist—to death, as it were—that "I believe that *p*" and "I desire that *q*" are taught—blindly—as alternative expressions of one's mind: they are ingrained as alternative ways

[34] So I agree with much Sydney Shoemaker (*The First-Person Perspective and Other Essays*, Cambridge, Cambridge University Press, 1996) has proposed but I place crucial emphasis on judgments and on the non-functionalist conception of beliefs (and desires) as commitments.

of expressing one's first-order beliefs and desires, other than asserting that *p*, or that *q* would be good to have.[35]

So rationality, or, equivalently, mental states as commitments, are necessary in order to perform the behavior which, in its turn, is necessary to be trained to use the relevant psychological vocabulary. That training consists in drilling someone to express her mind in a different way—namely, by asserting "I believe that *p*" (or "I desire that *q*") instead of merely asserting "*p*" (or "*q* would be good to have"). That suffices for acquiring the relevant psychological concepts and, once you have them, nothing else is needed to know your mind. Whenever you will be in a position to judge that *p* is the case (on the basis of evidence), or that *q* would be good to have, then you will also and immediately be in a position to avow (or to judge) that you believe/desire it. So, rationality and conceptual mastery—acquired in the way proposed—suffice, together, to give you knowledge of your own mental states for free—that is, without any cognitive endeavor.

The second half of the constitutive thesis: authority

But what about authority? Namely, how can we account for free, as it were, for the claim that when a sincere and conceptually competent subject avows her own mental states (or judges that she has them), she does have them? And even before engaging in this task, what grounds would there be to accept something as outrageous as the claim that any sincere psychological self-ascription made by a conceptually endowed subject is correct? Aren't cases of self-deception, however rare they might be, just a clear counterexample to that half of the constitutive thesis?

Well, good news, that half of the thesis is safe! Bad news, self-deception won't go away! Here's why. Akeel Bilgrami[36] has come up with an idea that I find illuminating: 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 mental state as a disposition. The irrationality is brought about by the clash between one's commitments and one's own unconscious dispositions. So take our jealous wife. 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 that he is faithful to her; yet she also has the unconscious belief, as a disposition, that he is unfaithful, which is operative in her inquisitive behavior. So, she is self-deceived all right, in the sense that she sincerely avows a belief and behaves in ways

[35] Indeed, this seems to me to be the right development of Wittgenstein's idea that avowals substitute behavior. It's just that when we move from avowals of sensations to avowals of propositional attitudes the behavior we must take into account is not merely physical but also linguistic.
For philosophers: I was pleased to find a similar point in Bar-On, *op. cit*., Ch. 7 *infra*: "It is important to see that the expressivist account of avowals proper does not turn on there being natural expressions of beliefs. Suppose there are no natural, non-linguistic expressions of occurrent beliefs. Still, we can appeal to the first-order linguistic expressions of beliefs as candidates for replacement by self-ascriptions (emphasis mine). The child says: "There's a cat" (looking at a rabbit) and her parent says: "You think that's a cat? It doesn't look like one" Next the child will learn to offer qualified expressions of thoughts by saying "I think that's a cat" and finally to express an occurrent thought by self-ascribing "I think Mom is going to give me a surprise!". However, Bar-On makes it in the context of articulating a purely neo-expressivist account of psychological avowals which explicitly rejects the constitutive model.

[36] See Bilgrami, *op. cit*.

that run contrary to it. Yet it is not the case that she has a false belief about her own beliefs. Rather she has two, different—both in kind and in content—beliefs that give rise to her distinctively irrational behavior.

Good. But then one may object that there are also cases of "negative" self-deception. Cases, that is, in which one says "I don't believe that *p*" yet behaves in ways that are explainable only by attributing to them the belief that *p*. Stretching the examples slightly,[37] but just because that would help make the point more vividly, think of Gianburrasca's brother in law—l'avvocato Maralli—"libero pensatore in città e bigotto in campagna"[38] or the always very wise Pascal who would say "I don't believe God exists (nor that he doesn't)" and yet would behave—or, at any rate, recommend to behave—as an irreprehensible Christian. In these cases subjects wouldn't be self-ascribing any belief. Hence, the only option seems to say that they falsely believe that they don't believe that God exists (nor that he doesn't).

But I think we can recast these examples in such a way that they cease to be a counterexample to authority. Here's how. We could say that the avowal is still the expression of subjects' mental states. Namely, of their commitment to not using "God exists" (nor its negation) as a premise of their practical and theoretical reasoning, which runs against the disposition to behave as kosher—if I may say so—Christians and thus use that belief as a premise of their practical (if not theoretical) reasoning.

So, having dispensed with one possible source of counterexamples let me then turn to the problem of explaining why authority holds. Remember, we want an account of authority that does not see it as the result of any cognitive achievement. For any cognitive achievement may go wrong and, therefore, there could be counterexamples to authority. But we have just seen that there aren't any.[39] So, the account must dispense with the result-of-a-cognitive-achievement picture, *tout court*.

The short answer is this:[40] if "I believe/desire that *p*" is issued in the way I have just described, then it will be judged on the basis of having judged "p" to be the case (or to be good to have) and of having thereby brought about the belief (or the desire) as a commitment that *p*. Hence, it is a priori true that any time one judges to have a certain mental state as a commitment, then one's judgment will be true. End of the story.

---

[37] These are cases of hypocritical judgments or behaviors and I don't want to convey the impression that self-deception should be thought of the same way. Characteristically, in fact, self-deceived subjects are simply unaware of having beliefs that run contrary to those they would explicitly avow. I'm using these examples just because they will help convey the kind of treatment that, *mutatis mutandis*, one should give of cases of negative self-deception.

[38] "A free-thinker in the city and a bigot in the countryside". See Luigi Vamba's *Il giornalino di Gianburrasca* arguably one of the best political and social satires of all times, though misguidedly portrayed as just a book for children, and one of the absolute masterpieces of Italian literature, regrettably not very well-known either abroad or in Italy (nowadays).

[39] This is not to say that one's own avowals are always correct but only that they are open to a very limited form of error: either they are incorrect because of conceptual incompetence (e.g. using "itch" for "pain" or *vice versa*) 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—if thought is anything over and above mental soliloquy).

[40] The longer answer—mostly for philosophers, I suppose, but feel free to read it if you wish—is in Appendix I.

So, let's recap the salient claims I've been making, which will be your take-home message:

(1) Self-knowledge on my view is not the result of any cognitive endeavor.
(2) What goes by the name of "self-knowledge"—that is, transparency and authority—is in fact guaranteed to hold a priori upon reflection on the concept of mental states as rational commitments and granted conceptual mastery.
(3) The key feature, which allows to avoid both observationalism and inferentialism, is the "blind" drilling that—I conjecture—is at the basis of our acquisition of psychological concepts.
(4) Mental states as commitments are normative—since they contain irreducible "oughts"—so it looks as if my final view on intentionality won't be reconcilable with a broadly naturalist perspective in philosophy of mind. But this, obviously, is left for further investigation.
(5) Similarly, it is left for further investigation how to stabilize the present view and how to extend it to all sorts of mental states.

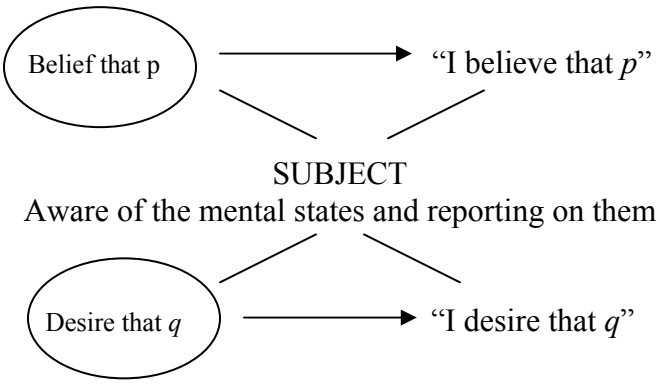
***A mo' di conclusione*: "Once upon a time in Eden"**

> Once upon a time we inhabited the Cartesian Eden. In the Cartesian Eden our souls were so pure that we could perfectly see through them. Anything going on in our minds was immediately and luminously presented to us and our reports on the events, which couldn't escape our inner gaze, were infallible.
> Then came the Austrian snake that made us eat from the tree of knowledge and we discovered that so much of what was going on in our souls was hidden from us. The transparency of the mind to itself seemed to be lost. We were no longer immediately in touch with our souls.
> The snake made us eat a little bit more and we saw that we could often be self-deceived. We could think a whole host of comforting thoughts and yet behave as if we did not believe them. Our reports on our mental states no longer appeared to be authoritative, let alone infallible.
> Having fallen from Eden we have been wandering in purgatory, thinking that we could never go back to were we thought we belonged.
> But the Cartesian paradise has, at last, re-opened its doors. Knowledge of our minds remains, as a matter of principle, more immediate and authoritative than knowledge of events in the outer world, or in other people's minds. It is just that it is less comprehensive than what we originally thought.

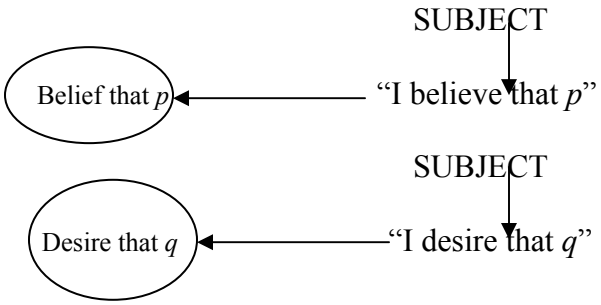
Appendix I

On Authority

The source of anyone's puzzlement when facing the problem of accounting for authority is—I believe—this. It seems that so long as one sees judgments and verbal psychological self-ascriptions as reports on, or descriptions of one's own mental states, the question arises of whether they are true or false and the temptation arises to think that if they are true, in fact always true, as we have just seen, it must be because they are true to the facts, facts which must be known to the subject one way or another. And we are back with the result-of-a-cognitive accomplishment picture. This traditional image can be visualized as follows:



The picture "says" that there is a subject who has the first-order mental state, is aware of it and judges that she has it. And the problem—to repeat—is that one could always be mistaken and, therefore, one's reports could always be wrong. But we have just seen that there is no reason to think that one could be wrong. So, a powerful corrective would be to invert the direction of fit and hold the following—constructivist[41]—picture:

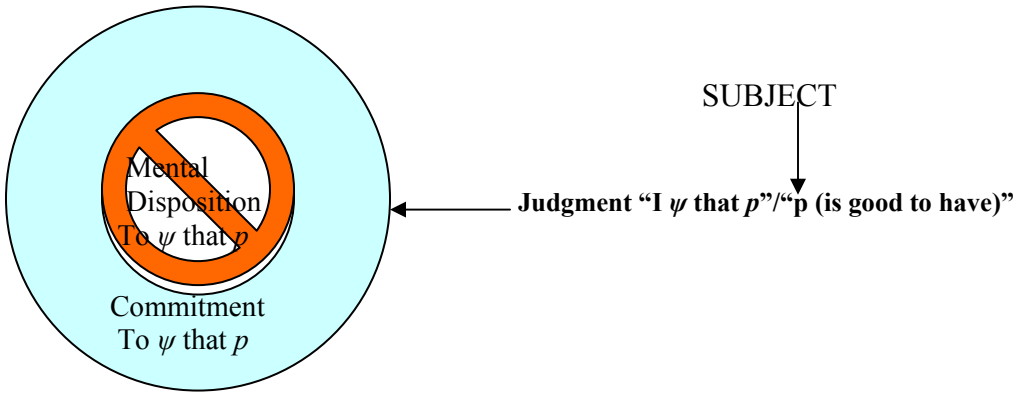


And the idea would be that it is the subject's judgment "I believe that *p*" or "I desire that *q*" which brings into existence the corresponding first-order mental states. On this model, there would be a sense in which it is absolutely true that we make up or create our minds. Moreover, since one's judgments would bring into existence the relevant first-order mental states, those judgments would be necessarily true. In fact, self-verifyingly so.

[41] This seems to be Wright's opinion in his early writings and Tanney's (*op. cit.*) as well.

Furthermore, there would be no temptation to think that one should know the first-order mental state in order to make one's judgment or avowal: 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 raise some eyebrows. In what sense do we create mental states? In virtue of what magical powers do we do that? How could saying "I believe that *p*" suffice to bring about the corresponding first-order mental state? 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 mental states. For mental 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 would be good to have) we do create our beliefs and desires as commitments. Moreover, "I believe that *p*" and "I desire that *q*", when used as alternative means to express one's mind, are necessarily true, since they are made true by those mental states as commitments which have been brought about by those judgments they give expression to. Finally, when viewed this way, it becomes less mysterious how actually judging or sincerely asserting "I believe that *p*", or "I desire that *q*" could bring about the corresponding first-order mental states. For once we are conversant with the practice of substituting "I believe that *p*" or "I desire that *q*" to the assertion of "*p*" and "*q* would be good to have" we can bring about the relevant first-order mental states either by judging that *p* is the case (or that *q* would be good to have) or by judging that one believes that *p* or desires that *q*. Let's visualize what's being proposed here:



So, when used in the way proposed, judgments (or sincere assertions) like "I believe that *p*", or "I desire that *q*" are like performatives, namely like "I promise to pay you $20", "I thee wed", "I hereby name you so-and-so". They do make certain things happen, for they do create the first-order mental states as commitments.[42]

[42] Footnote for philosophers. Still, they retain truth-evaluable content and are also reports on what has being done, viz. committing oneself to either holding p true, or that it would be a good thing to have. To see this, consider that performatives (and their mental counterparts, viz. judgments) are sentences which absolve

So, to conclude and recap:

(1) "I believe that *p*"/"I desire that *q*" are true, in fact always true, because they are alternative means to express those mental states (as commitments) that have been created by the subject by judging that *p* or that *q* is good to have.
(2) They themselves can have, besides an expressive function, a performative one.
(3) The use of the word "knowledge" in the locution "self-knowledge" is appropriate only insofar as it is taken to be a "grammatical" use that is signaling the fact that we cannot fail to know what our mental states are: if psychological self-ascriptions are made by being in the psychological states they manifest and (can themselves) bring about, while making them we cannot fail to know which mental states we are in. But in no way does the use of that word signal the fact that self-ascriptions of mental states are the result of any cognitive achievement.

Finally, to state our answer to the issues at the heart of the debate among constitutivists:

(i) the grounds for the constitutive thesis are rationality—understood along the lines which have been proposed here—and conceptual mastery—which, crucially, has to be seen as depending on blind drilling.
(ii) The constitutive thesis has a slight anti-realist flavor for, although not all mental states are created by the subject, those as commitments are.[43] This, however, is not to be confused with the irrealist claim that mental states don't really exist.

---

more than one function at the time: they make things happen but they also say what is being done by means of them. Hence, they have truth-evaluable content. For instance, "I promise to give you 20$" is true iff AC promises to give you 20$ and it could be false, since I could be lying. Similarly, "I believe it is raining" is true iff AC believes that it is raining and it could be false, since I could be lying. They also admit of negations that don't express a lack of commitment but the presence of "negative" commitments (which is not yet a commitment to the contrary). So, for instance, "I don't believe that *p*" (when it isn't a notational variation of "I believe that not-*p*") means that I commit myself to not using *p* as a premise for my reasoning and deliberating procedures. Furthermore, performatives can be embedded in wider contexts, such as suppositions and, therefore, conditionals. "Suppose I believe that *p*" is thus to be understood as "Suppose I commit myself to *p*'s truth, then here are the consequences".

Therefore, I completely agree with Bar-On, *op. cit*., that expressivist construals, once properly formulated, don't fall pray of Geach's objection. But, ultimately, I disagree with her on a number of fronts: in particular, that a constructivist reading of mental states necessarily entails metaphysical anti-realism (or irrealism) about mental states and that expressivism is compatible with genuine knowledge. For, as to the former complaint, it seems to me that the fact that the existence of certain mental states is partly dependent on our judgment does not make those mental states less real. In other words, judgment-dependence is a claim about the provenance of those mental states, not about their (ir-)reality. And, as to the latter, it seems to me that the use of the word "knowledge" in the locution "self-knowledge" is appropriate only insofar as it is taken to be a "grammatical" use that is signaling the fact that we cannot fail to know what our mental states are: if psychological self-ascriptions are made by being in the psychological states they manifest and (can themselves) bring about, while making them we cannot fail to know which mental states we are in. But in no way can the first-order mental state constitute a subject's reason for her avowal (for, roughly, this will bring us back to the observational model) nor can the avowal amount to knowledge because it is reliably produced by a suitable cognitive mechanism (for the reasons reviewed in the section on Armstrong). Since *tertium non datur*, we'd better leave it as that and deflate the use of "knowledge" in the locution "self-knowledge" and strip it of all its usual cognitive flavor.

[43] Notice, moreover, that on my view there is just one mental state (as a commitment)—the first-order one—that, when conceptualized in the ways proposed, can play a broader functional role. Thus, for

Appendix II
The inadmissible question—"What's the point?"

This section is only for unsympathetic readers—those who would feel like asking "What's the point?". The sympathetic readers, if there are any, should just skip it.

At the end of a talk like this to an audience mainly composed of non-philosophers the most likely reaction one gets is of annoyance, boredom, and incredulity—or so I fear. Most people just walk away keeping these thoughts to themselves: a mixture of laziness, good manners and sheer indifference will make them slip into silence during the discussion period and then walk out of the room as soon as possible. Well, I won't let this happen. I tell you in a minute what you should do instead. But, before telling you, let me remind you of the following.

First, what's the point of Michelangelo's frescos in the Sistine Chapel? None, whatever. Or, for that matter, what's the point of Dante's *Divine Comedy*, or of playing chess, or of playing any sport? None, whatever. But the world would be a much duller place without those works of art, those spectacularly intelligent games and the unforgettable movements of Michael Jordan, Maradona—when he was still in relatively good shape—or of John MacEnroe—do you remember the beauty of his serve-and-volley game?! Similarly, the world would be a much duller place if there wasn't room for philosophical reflection. It has its own kind of beauty.

Secondly, we have been defined as rational animals (although, sometimes, I wonder whether the definition applies…). Hence, to engage in rational thought is, at least in part, a fulfillment of our human nature. Philosophy is, first and foremost, rational inquiry, and, therefore, a way of fulfilling our nature.

Finally, it is obviously important to try and understand something about one of the features, which seem so fundamental to what we are. You often hear people draw the line between human beings and brutes—animals—by appealing to self-consciousness. Well, a large part of what self-consciousness consists in is precisely the distinctively immediate awareness each of us has of her own mental states. Surely many other animals move about in their environment in purposive ways displaying intelligent behavior, and there is no reason to think that they don't have mental states. And it is equally intuitive that they have a sense of themselves as bodily entities distinct from other objects and animate beings in their environment. However, it's more difficult to think that they are also aware

---

instance, if I have the concept of belief I can perform the following inference: judge that *p* is the case and, therefore, that there is something I believe, which I couldn't have performed without having that concept, even if I could already have had that belief as a commitment. So, as a matter of fact, there is no need for copying mechanisms of the kind envisaged by Armstrong. Mother Nature has given us all we need in order to have knowledge of our own mental states by giving us the ability to make judgments, rationality and conceptual capacities. So, I perfectly agree with the spirit of Shoemaker's famous remark: "From an evolutionary perspective it would certainly be bizarre to suppose that, having endowed creatures with everything necessary to give them a certain useful behavioral repertoire—namely that of creatures with normal intelligence, rationality, and conceptual capacity, plus the ability to acquire first order beliefs about the environment from sense-perception—Mother Nature went to the trouble of installing in them an *additional* mechanism, a faculty of Inner Sense, whose impact on behavior is completely redundant, since its behavioral effects are ones that would occur anyhow as the result of the initial endowment". (Shoemaker, *op. cit*., pp. 239-40).

of their own mental states. If true, this would make life much easier for them, in some ways, but, also, much duller: surely they can't assess their own thoughts, not having access to them; nor, by the same token, can they morally evaluate them. This will certainly spare them the trouble of trying to get them right, or of being morally reproachable and of possibly undergoing punishment. But they won't ever feel the pleasure of being right either, or of feeling elated at the thought of having done the right thing. An easy life is very often a boring life. A life that not even in moments of despair we should feel inclined to.

So, don't ask "What's the point?"—you know what the point is, don't you?! Now I tell you what you should do instead: go back to your notes and ask anything, anything which doesn't seem clear, or that doesn't convince you and try to articulate your reasons. In brief, play this game, and you'll see where the fun is.