**Self-Knowing Interpreters**

Annalisa Coliva

Philosophy Department, University of California Irvine

Professor Annalisa Coliva

Department of Philosophy

85 Humanities Instructional Building

University of California

Irvine, CA 92697-4555

(949) 824-6525

**email**: annalisa.coliva@unimore.it

In this paper, I first present the characteristic features of what may be regarded as truly first-personal self-knowledge. Namely, groundlessness, transparency and authority (§1). The key idea defended in the following is that they are necessary and a priori aspects of what goes by the name of “(first-personal) self-knowledge”. For massive failures at self-knowledge would display either the lack of the relevant psychological concepts or failures at rationality, understood in a “thick” sense, which is accordingly specified. These characteristic traits of first-personal self-knowledge are then defended against possible objections stemming from skepticism regarding self-knowledge deriving from recent findings in cognitive sciences (§2). It is argued that none of this shows that we never have essentially first-personal self-knowledge. Rather, it shows that its scope is limited and does not extend to our deep seated and future dispositions, to the dispositional elements of our feelings and emotions, and to the causal relations among our various mental states, which are known, if and when they are, in a third-personal way. Various modes of third-personal self-knowledge are then presented and discussed (§3). Yet, all this is compatible with the fact that we have essentially first-personal knowledge of a wide range of mental states.

1. **First-personal self-knowledge**

First-personal self-knowledge, which consists in the ability correctly to self-ascribe mental states, like “I am in pain”, “I see a canary”, “I believe it will rain tomorrow”, “I am scared of that dog”, is puzzling, from an epistemological point of view. For, on the one hand, it regards contingent states of affairs. On the other, however, it does not seem to be obtained by the same means as the ones we employ when we form knowledge regarding contingent states of affairs concerning the physical world or other minds. In particular, first-personal self-knowledge seems neither inferential nor observational. For, of course, we sometimes find out our own mental states by means of inference starting with the observation of our behvior and further mental states. However, for the inference to get started at all, we must already have knowledge of those mental states of ours that figure as contents of the premises of our reasoning. On pain of an infinite regress, there must be knowledge of our own mental states that is not inferential. Somewhere down the line, we must be able to know our own mental states directly.[[1]](#footnote-1)

Moreover, quite intuitively, first-personal self-knowledge cannot be a matter of observation either: first, mental states are just not the kind of thing one could observe. 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. Very briefly: it is 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 go—how would I know that other people have mental states at all, if those mental states are intrinsically private to them and foreclosed to me? It may really become a recipe for catastrophe when taken to involve conceiving of our psychological language as private. Since Wittgenstein (1953), however, private language has become synonymous with 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. Still, if the meanings of our 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. 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. Yet, as we have just seen, that distinction is essential to there being a (psychological) language at all. Thus, Cartesianism about self-knowledge may take us to a conception of our psychological language, which turns it into a non-language. However, we do have a psychological language—after all we tell each other about our own thoughts and feelings, sensations and emotions all the time and we seem to understand each other well. Therefore, it seems safe to conclude that Cartesianism will not do. If Cartesianism has to go, so has the idea that self-knowledge could be observational.

First-personal self-knowledge is therefore groundless. That is to say, it is based

(1) neither on observation

(2) nor on inference.

If groudlessness so characterized consists in the idea that (first-personal) self-knowledge is not the result of any substantial cognitive achievement, such as observing or inferring from a symptom to its likely cause, it does not necessarily involve the idea that one’s psychological self-ascriptions are not based on anything, such as the very experience one is undergoing when, for instance, one is in pain and avows it.

That said, it has to be registered that many philosophers think that the on-going experience, which would allegedly ground one’s psychological self-ascription, would not play the role of genuine evidence, or of a reason, for one’s self-ascription. Accordingly, self-ascribing pain, while undergoing a painful sensation, would simply amount to recognizing what is pain as pain, or even to giving immediate expression to one’s on-going pain through conceptual and/or linguistic material. It would not amount to having an independent mental state in view which could, as such, corroborate the truth of one’s self-ascription. For having that mental state “in view”, in such a way as to ground one’s self-ascription, would, presumably, be identical to, or at least extremely close to, already recognizing it as the mental state it is. Still, recognizing something for pain is just—or is very close to—judging “I am in pain”. The justification of the relevant self-ascription that should be underwritten by the mental state would thus be either straightforwardly circular or based on a mental state too tightly connected to the self-ascription one would like to justify thereby. These theorists would therefore subscribe to a stronger notion of groundlessness, according to which mental self-ascriptions, characteristic of first-personal self-knowledge, not only fail to be the product of observation and inference, but are not based—in any epistemologically relevant way—on one’s previous awareness of on-going mental states.

**Groundlessnessweak**: first-personal self-knowledge is neither observational nor inferential.

**Groundlessnessstrong**: first-personal self-knowledge is neither observational nor inferential, nor is it epistemologically based on one’s previous awareness of one’s on-going mental states.

For our purposes, it is enough to register these different conceptions and to point out that, in a minimal sense, groundlessness amounts to the idea that first-personal self-knowledge is not the result of any substantial cognitive achievement such as inner perception or inference to the best explanation.

Now, suppose that a subject, endowed with the relevant concepts, and quite proficient in applying them to herself and others, did so, in her own case, by *always* observing her behavior and inferring to its likely cause. She would not know that she is in pain because she feels it, but because she would see herself scream and moan, after getting injured, and would infer that she must be in pain. She would not know that she intends to do some groceries because she is aware of that intention, but because she would realize that she has gone to the shops and done her shopping. Such a subject would strike us as someone incapable of having genuine first-personal self-knowledge and in some sense as alienated from herself. The word “alienation” may sound grandiose in this connection, but, minimally, it would mean that the subject would bear to herself—that is, to her mental states—the kind of relation she bears to those of others. However, what makes a mental state of ours, ours is, beside its origin in our brain activity, the fact that it is known to us in a way which differs from how we know of other people’s mental states, and in which no one else can know of it, so as to figure among the subject’s motivations for her further actions. The failure of groundlessness would therefore result in a lack of “ownership”, or “miness”, as we might put it, over the mental states one would be inferring to have. Or, again, in the case of sensations and other mental states with a distinctive phenomenology, not knowing them directly, but by having to infer to them, starting from the observation of one’s overt behavior, by application of a little theory, therefore hypothesizing having them, would cast doubt on the fact that one is capable of enjoying sensations and other phenomenologically salient mental states at all. Moreover, when intentions (and other propositional attitudes relevant to action) are at stake, a subject who would know of them only inferentially, from reflection on her behavior and by inference to its best explanation, besides being “alienated” from her actions, would actually strike us as not fully responsible for them. For she would know of her motives only afterwards, after having witnessed herself act in certain ways. Yet, at the time when the action was accomplished she would not know either what action she was in fact doing, or why. She would not know her action, because, prior to having interpreted it, it would just be a movement or a piece of overt behavior to her. Moreover, failing to have immediate access to her intentions, she would not know why, for instance, she happens to be pushing a cart around in a supermarket—as she witnesses herself doing. She would not act but be acted and at any instance of the question “Why did you do/are you doing that?” she would be entitled to respond “I do not know. Let me find out” and she would then be entitled to preface whatever answer she might wish to give, based on inference to the best explanation, with “Probably I did/am doing that because…” or “It is my hypothesis that I did/am doing this because…”. Now, these responses, though understandable from the point of view of a subject like the one just described, would actually strike us extremely odd. They would strike us as reports on another subject, who is the subject of one’s actions, which would happen to live in one’s body.[[2]](#footnote-2) On a “thick” notion of being a rational agent, who can be considered rationally responsible for (at least the vast majority of) her actions, knowing of them and their motivations only through inference to the best explanation and in the way of a hypothesis would actually represent a major departure from rationality.[[3]](#footnote-3) Hence, groundlessness seems to be an element of genuine first-personal self-knowledge, which is in turn a constitutive element of being a responsible agent.

To claim that groundlessnessweak is constitutive of first-personal self-knowledge and to redeem that claim by means of a priori considerations having to do, as we have seen, with the notion of sensation (and possibly other phenomenologically salient mental states), and with that of being rational agents who can be held responsible for their actions, does not mean to prejudge the issue of how a satisfactory account of first-personal self-knowledge can go about trying to meet this desideratum. In particular, it is still possible to propose a fully epistemic account of first-personal self-knowledge that aims to respect this requirement. It is only by the lights of non-epistemic theories of first-personal self-knowledge that groundlessness, just like the other features of first-personal self-knowledge we will examine in the following, cannot be satisfied by epistemic accounts and has to be considered an a priori feature of what goes by the name of first-personal self-knowledge, which is not underwritten by any epistemic achievement on a subject’s part, but can only be redeemed on the basis of a priori reflection on a set of mutually interdependent concepts. We will not have to address this issue here. Still, it is worth bearing in mind that defending groundlessness on a priori grounds is compatible with different accounts of first-personal self-knowledge.

Consider a sensation of cold or the deliberation to cook risotto for dinner. If someone asked you “Are you feeling cold?” or “What do you intend to cook for dinner?”, you would be in a position to answer their question immediately. Our mental states seem to be directly, or transparently known to us. More precisely, their occurrence is usually of a piece with one’s awareness of them. According to some theorists (Shoemaker 1996), they are self-intimating. If one has them one would be *ipso facto* aware of them. Similarly, the occurrent painful sensation, or fear one might have is of a piece with one’s awareness of them. Again, one’s seeing a PC in front of one is of a piece with one’s awareness of it. Surely, however, the occurrence of states of affairs out there—even the most banal ones—is not of a piece with one’s awareness of them. The trivial fact that there is some pigeon flapping its wings right now outside the window, for instance, is not of a piece with my awareness of it. I will have to look out of the window and recognize the bird as a pigeon and so on. Nor is the occurrence of other people’s mental states of a piece with one’s awareness of them. Hence, for example, no matter how good one might be at figuring out what is crossing someone else’s mind right now, that someone else is feeling bored, perplexed or annoyed by what they are reading is not something one can be immediately aware of. As said, it is something one will have to *figure out* by taking into account facial expressions, sighs, bodily movements, connect them with general knowledge of what those reactions are an expression of and finally infer that they are bored, perplexed or annoyed. Transparency, then, is one of the features that set first-personal self-knowledge apart from all other kinds of knowledge.

Some theorists are unconvinced that transparency is a characteristic feature of our own mental states (see Snowdon 2012 for a criticism of transparency with respect to sensations, emotions, and passing thoughts). They point out, for instance, that higher-order mammals and infants do not seem to be capable of self-ascribing sensations and basic emotions, or even perceptions, while they can enjoy them. To such a worry, one might respond by noticing that we have been talking of transparency mostly in terms of awareness and it seems safe to hold that at least in the case of sensations, occurrent emotions with a characteristic on-going phenomenology, and conscious perceptions, their occurrence seems to coincide with one’s awareness of them, even if one does not possess the concepts necessary to self-ascribe them. More precisely:

**Transparencyweak**: if one has a given mental state M, one is aware of it. That is to say, the mental state M is phenomenologically salient to the subject.

Even so, it is clear that transparency does not hold unconditionally, but only for those mental states which have a distinctive phenomenology—leaving aside for the moment the width of this class—and are not purely dispositional. Furthermore, subjects undergoing these mental states will have to be cognitively lucid, attentive and alert.

However, one might object to this characterization by pointing out that it captures self-awareness rather than self-knowledge. For the latter, contrary to the former, consists in the ability correctly to make the relevant psychological self-ascriptions. Furthermore, one might want to capture the idea that also propositional attitudes that may lack a distinctive phenomenology can be transparently self-known. As we saw at the very beginning, having formed the intention to cook risotto for dinner, one may be immediately in a position to self-ascribe that mental state, even if, arguably, there is no distinctive phenomenology to forming intentions. In such a scenario, the likely move would be to conditionalize transparency to the obtaining of the relevant (C-)conditions. Namely, subjects should be conceptually endowed, besides being cognitively lucid, attentive and alert. By means of such a conditionalization, then, usual counterexamples to transparency, often presented in the literature on self-knowledge, would easily be dispensed with. The case of animals and infants would be countered by the requirement that subjects should possess the relevant concepts. If, in addition, a subject is under the effect of drugs, has her attention occupied by a pressing task, or is distracted, she may be having sensations, emotions, perceptions, and yet not be aware of them. Given the conditionalization just proposed, these cases would not represent counter-examples to transparency.

A separate remark is apposite in relation to the Freudian idea that we may have many unconscious propositional attitudes, such as beliefs, desires, and intentions, of which, by definition, we are totally unware. Obviously, there is no denying that this might be the case. However, Freudian mental states are mental dispositions. Hence, one may add to the C-conditions that when propositional attitudes are at stake, they should be beliefs, desires and intentions as commitments, not as dispositions. I cannot go into details here (but see Coliva 2015 and 2016). The basic distinction is that propositional attitudes as commitments are normatively constrained and in fact necessarily self-known, while dispositions are not. Hence, transparency can be characterized as follows.

**Transparencystrong**: Given C-Conditions (including concepts’ possession, cognitive well-functioning, alertness and attentiveness and to the exclusion of unconscious and purely dispositional mental states) if one has a given mental state M, one will be in a position to judge and/or believe that one has it.

Now we should consider whether transparency is a contingent feature of first-personal self-knowledge or, rather, a necessary and a priori aspect of it. Let us start with sensations, perceptions and basic emotions. One may distinguish between sensory states, with no representational content and therefore without correctness conditions, which simply register proximal stimuli, and sensations properly so regarded. Only the latter are necessarily conscious—that is, such that having them coincides with being aware of them. That is to say, transparency is, on this reading, a necessary and a priori aspect of what goes by the name of “sensation”. Similar considerations may be put forward in the case of at least those emotions that have a distinctive phenomenology to them. However, the claim that transparency holds a priori, at least for some kinds of mental state, would then seem to depend on a terminological choice. Furthermore, one may want to insist that for instance the soldier who has been marching for days does feel pain, even if he is not aware of it because his fatigue and stress are so conspicuous that he is oblivious to them and just keeps walking. It is at this stage that the C-conditions we have introduced in the characterization of transparencystrong are helpful to delimit those cases in which it makes sense to think that our knowledge of our on-going sensations meets the transparency requirement. If a subject were always unable, while feeling pain and being attentive and alert, to judge that she is, this would cast doubt on the fact that she does possess the relevant concepts; alternatively, if we had reasons to think that she does have those concepts, her inability to self-ascribe an on-going pain would cast doubt or on her cognitive well-functioning to the point of impairing the idea that we are dealing with a rational subject. Furthermore, she would not be in a position to take responsibility for her actions. She would try to cure her wound, for instance. If asked why, however, she would not know, even after screaming and moaning and while having, *ex hypothesi*, the relevant psychological concepts.

The case of perceptions is clearly different, for, alongside with conscious perceptions, whose occurrence would be of a piece with one’s awareness of them, we have unconscious ones. Blind-sight would be a case in point. Hence, there would be perceptions whose occurrence would not be of a piece with one’s awareness of them. To such a challenge, we can respond by noticing that clearly unconscious perceptions fall out of first-personal self-knowledge. For we would know of them by observation and inference, by having witnessed ourselves act in ways which can be made sense of only by presupposing that we did have those perceptions. By contrast, if we were dealing with conscious perceptions, then, their occurrence would be of a piece with one’s awareness of them. Furthermore, if endowed with the relevant concepts, while being cognitively alert and attentive, we would be in a position immediately to judge (or believe) that we are having them.

It is worth noting that, given the strong link between perception and action, the actions we would perform based on unconscious perceptions would fall out of responsible agency. The subject affected by blind-sight does not know either that she is catching a ball as she does that or why. Hence, she cannot be held responsible for it. Transparency is therefore a necessary and a priori element only of our knowledge of conscious perceptions, which are, in turn, constitutive of responsible outer actions.

Similarly, a subject who, having formed an intention or other propositional attitudes as commitments, were systematically unable to self-ascribe them, would either be lacking the relevant concepts; or else, if there were reasons to think she possesses them, she would seem to be somehow mentally deranged. For she would systematically respond “I do not know” to the question “Do you believe/desire that P/intend to φ?”, after in fact having asserted (or judged) “P”, “P is worth-doing” and “I will φ”, while having the relevant psychological concepts.

Once more, claiming that transparency is a constitutive feature of first-personal self-knowledge and that it can be redeemed on the basis of a priori considerations concerning a range of interconnected concepts, such as the concept of responsible agency, rationality and first-personal self-knowledge, does not preclude the possibility of giving an epistemic account of it. It is only by the lights of non-epistemic accounts of first-personal self-knowledge that this requirement is not underwritten by any kind of epistemic performance on a subject’s part and is in fact just a conceptual truth concerning a set of interlocking concepts. Once again, we do not have to address this issue here, but it is worth keeping in mind that defending transparency on a priori grounds is compatible with different accounts of first-personal self-knowledge.

Let us now consider authority. 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 that summers in Italy are really too hot” and you are sincere and know how to use ‘I’, ‘think’, ‘summers’, ‘Italy’, …, then nobody could challenge you by saying “Are you sure that this is what you are thinking?”. Of course, one can challenge the subject’s grounds for believing that summers in Italy are too hot, by pointing out that it is August 15, you are in Italy and there are only 15 °C. Yet, one cannot challenge the subject’s own believing that she believes it. By challenging the grounds in favor of that content’s belief, one can lead a subject to revise her belief, but this possibility depends on the (implicit) acknowledgement that the subject does believe—erroneously, from your point of view—that summers in Italy are too hot and that she knows that much about herself. Similarly, if I sincerely avow a painful sensation, it would be pointless for someone to challenge me by saying “Are you sure that is what you are feeling?”. To the restatement of my avowal I could only add that I am being sincere and that I can speak English. Therefore, subjects are authoritative with respect to their own mental states: if they say (or judge) that they have a certain mental state M, then they have it.

Surely, however, if someone asked you “What is the weather like?” or “Is Elly sad since her sudden loss?”, from your sincere and conceptually competent answers it wouldn’t follow at all that it would be inappropriate for someone to challenge your claims. Therefore, authority is another feature that sets self-knowledge apart from all other kinds of knowledge.

Again, some theorists are wary of authority (see again Snowdon 2012). With respect to sensations, for instance, they point out that one can envisage a situation where a subject is asked to locate her sensation and is uncertain about that. This, however, would merely show that we are not authoritative with respect to the bodily location of our on-going sensations. Yet, it would not show that we are not authoritative with respect to the fact that we are undergoing them.

Another case often discussed in the literature is the one of a subject who, by dreading the pain a dentist will produce by messing about her mouth, claims to be feeling pain already when the dentist has not even touched her. This case can be taken care of by specifying the C-conditions under which authority is supposed to hold. In fact, the subject should be cognitively lucid. Arguably, however, intense fear can impair one’s cognitive functions. Once again, the possibility, on specific occasions, of mistaken self-ascriptions of sensations, should not be taken as a counterexample to authority, but only as imposing the need of properly characterising its extent.

Another prima facie difficult case to handle is the one of confused or vague sensations, such as itches that border pain. A subject could then be unclear whether she is undergoing one or the other kind of sensation and make avowals she would then like to retract. This, however, would not be a counterexample to authority, because we are in fact dealing with a confused or vague sensation, which escapes definite categorization. Hence, we vacillate between two concepts we do possess, when both of them seem to apply to it to some extent. In this sense, our self-ascriptions may be imprecise and retractable. Yet, this does not show that, at least when sensations are not of a confused nature, we would not be authoritative with respect to them. More generally, the ubiquitous presence of vagueness in our concepts should not be taken—by itself—as a serious challenge to the possibility of first-personal self-knowledge. No more than the ubiquitous presence of vagueness in the domain of colour is normally taken as a challenge to the possibility of correct colour perceptions and judgments.

Another possibility is to be wrong about the content of one’s perceptual appearances. For instance, one can say that one’s after image contains five red dots, when in fact it contains only four. However, authority can be maintained even in this case because determining the number of dots involves counting and one may go astray in doing it. Still, one would be authoritative with respect to the fact of being having an after image.

Interestingly, Eric Schwitzgebel (2008) has recently argued that we are not authoritative with respect to our inner feelings and emotions. The idea is that we may be bad at judging that the feelings we are currently experiencing are feelings of anger, say, when in fact this would be clear to a third party. It should be admitted that sometimes we do not realize what the feelings we may experience are *symptoms* of. This, however, could be due to poor performance in connecting one’s on-going feelings to the dispositional elements of the relevant emotion. Hence, it would be a case of mistaken self-interpretation. That is to say, it would be a case of lacking or of going astray in one’s third-personal knowledge of one’s own mental states, rather than a case of failure at first-personal knowledge of them.

Finally, moving on to propositional attitudes, it may happen that one says that P, or that P would be good to have, thus manifesting the belief or the desire that P, and yet behaves in ways that run contrary to one’s professed beliefs and desires. This phenomenon, known as self-deception, is taken to be the most powerful counterexample to authority. There are three possible answers on behalf of the supporters of authority. First, one may complicate the C-conditions to exclude cases of self-deception from the range of propositional attitudes one is authoritative about. This move, however, seems quite *ad hoc*, hence it would be preferable to avoid it. Second, one may propose an alternative account of self-deception, which is compatible with the retention of authority over one’s propositional attitudes (see Bilgrami 2006 and Coliva 2016). Finally, one may hold that authority can admit of exceptions, yet argue that it is present in a significant amount of self-ascriptions about one’s own intentional mental states.

Hence, we face the possibility of two different readings of authority, which differ in strength. A minimal one has it that we are mostly—though not always—authoritative with respect to a specified class of mental states we can enjoy, and, in particular, that self-deception may lead us to wrong psychological self-ascriptions. A stronger reading, in contrast, has it that, as a matter of fact, there are no relevant exceptions to authority, once we have appropriately specified the class of mental states for which it holds and have independently accounted for self-deception in ways which explain away the impression that it should pose a challenge to authority. Accordingly, we would have the following two possible readings of authority:

**Authorityweak**: Given C-Conditions (including concepts’ possession, cognitive well-functioning, alertness and attentiveness) if one judges to have a mental state M (save for dispositional ones or for the dispositional elements of some mental states), one will usually have it.

**Authoritystrong**: Given C-Conditions (including concepts’ possession, cognitive well-functioning, alertness and attentiveness) if one judges to have a mental state M (save for dispositional ones or for the dispositional elements of some mental states) one will always have it.

Be that as it may, authority too would seem to be a necessary and a priori feature of first-personal self-knowledge. For, if one were systematically proved wrong in one’s psychological self-ascriptions, doubt would be cast upon one’s possession of the relevant concepts. Alternatively, if one wished to maintain that a constantly mistaken subject (with respect to her own mental states) could still be said to have the relevant psychological concepts, perhaps because she retains the ability to apply them to other people’s mental states, her persistent self-deception would impair her rationality. For she would avow certain mental states, while she would systematically behave in ways which run contrary to them. Hence, we could no longer make sense of her linguistic and non-linguistic behavior.

Once again, to claim that authority is constitutive of first-personal self-knowledge and that it is a priori connected with other notions, such as the one of being a rational agent who can be held responsible for her actions, does not preclude the possibility of accounting for it in epistemic terms. It is only in the perspective of non-epistemic theorists that those attempts fail and that authority is not underwritten by any epistemic achievement on a subject’s part but is purely a fall-out of the necessary and a priori links holding between some interconnected concepts.

Groundlessness, transparency and authority, therefore, are constitutive features of first-personal self-knowledge, and are constitutively tied to what it means to possess psychological concepts, as well as to what it means for someone to be a rational subject, who can be held responsible for her own actions.[[4]](#footnote-4)

1. **Counterexamples from cognitive science?**

Recent studies in cognitive psychology tend to challenge the idea that we are authoritative with respect to our own mental states. In what follows, I will draw extensively on Brie Gertler’s (2011) excellent discussion of the topic.[[5]](#footnote-5) First, there are studies that impugn the idea that we have privileged access to our character traits, such as jealousy or loyalty. People tend to deny being jealous and to affirm to be loyal, even if they actually behave in ways which run contrary to their psychological self-ascriptions. Second, it has been shown that we are bad at “affective forecasting”.[[6]](#footnote-6) That is to say, we predict that a life-changing event, such as winning the lottery, will permanently affect our psychological attitude towards life. Studies have shown, however, that after a short period in which one’s moods are elated, people revert to their “happiness baseline”. In both cases, we may have privileged access to our resolutions or intentions and expectations, such as the intention to be loyal, or the expectation that winning the lottery will make us permanently happy, and yet we are not authoritative with respect to whether we will actually behave in the way we intend or expect we will behave. Gertler rightly notices that both character traits and affective forecasting involve access to our own psychological dispositions. Dispositions involve counterfactual conditions and it may well be the case that we are bad at determining whether we will satisfy them or not. This clearly limits the scope of first-personal knowledge of our own mental states, but does not rule it out.

Other studies show that we are bad not only at determining our psychological dispositions, but also at indicating the causes or motivations of our actions.[[7]](#footnote-7) In fact, they often remain unconscious, by operating at the subpersonal level. Hence, we do have first-personal knowledge of certain occurrent thoughts, such as one’s desire to have ice cream, which we take to be the cause of our action of going towards the fridge to have some, but in fact that thought may not be the real cause of our action. Again, Gertler rightly notices that, since Hume, causes have been shown not to be directly observable and rather be the result of theorizing about established correlations between different types of events. Therefore, it should not be surprising that we lack authoritative access to the causes of our actions. Still, we are under the illusion of having a distinctively first-personal access to them because we do have such privileged access to the thoughts that we take to be the causes of our actions. Once more, the scope of first-personal knowledge of our own mental states appears to be limited, but not to be vanishing.

Similar results have been achieved by studies concerning our moods, such as pessimism, like the tendency to regard the future as bleak, and emotions, such as anger. Gertler correctly points out that moods are dispositional mental states and that emotions involve, besides a certain feeling, with respect to which we do have first-personal access, certain dispositional elements. This idea chimes well with the borderline conception of the emotions defended in Coliva (2016), according to which emotions cannot be reduced to either feelings or dispositions. It may well be, then, that while we have privileged access to their phenomenological aspects, we do not have it to their dispositional elements.

Finally, if we do not have first-personal knowledge of dispositional mental states, that entails that we are not authoritative, nor do we have immediate and groundless access to a lot of propositional attitudes, such as beliefs and desires, which are dispositional. Indeed, this is only to be expected. Yet we do also have the impression of having first-personal access to our beliefs, desires and intentions. Our view, presented in chapter one, that we should distinguish between different kinds of propositional attitudes—that is, those as dispositions and those as commitments—comes in handy at this stage. For we may say that while we lack first-personal knowledge of the former, we have it of the latter. Yet, while this imposes a further limitation to first-personal self-knowledge, it does not show that we lack it altogether. Rather, it merely shows that first-personal self-knowledge is limited to occurrent sensations, including those characteristic of at least basic emotions, to perceptions, current thoughts and propositional attitudes as commitments.

1. **Third-personal self-knowledge**

So far, we have examined the characteristic aspects of first-personal self-knowledge. That is to say, those aspects, which set our knowledge of our own mental states apart from other kinds of knowledge, such as knowledge of truths about the external world, usually achieved through perception, or about other people’s mental states, customarily obtained through observation and inference to the best explanation, starting from the observation of their overt behavior. The various limitations to first-personal self-knowledge we considered, however, show that ample room for error is provided in figuring out our own mental states. This can be explained by noticing that we often acquire knowledge of our mental states in a third-personal way, even in our own case, as a result of observation of our behavior and inference to the best explanation; or else, thanks to inference to the best explanation starting from one’s first-personal knowledge of one’s on-going sensations, feelings, perceptions and propositional attitudes. We therefore possess not only first-personal but also third-personal self-knowledge. In general, that is, we are often self-knowing interpreters, as we may put it.

When third-personal self-knowledge is at stake, however, none of the characteristic features of first-personal self-knowledge is present. There is no groundlessness, for indeed our knowledge of our own mental states is either based on observing our behavior and inferring to its likely cause, or it is based on our awareness of our inner phenomenology and to inferring to its probable explanation. Two avenues for error are open, though. For we may mischaracterize our own behavior and hence start out our inferences with mistaken premises. Or else, we can be wrong in identifying a certain mental state as the likely cause of our on-going feelings or overt behavior. Thus, authority does not hold either. Nor does transparency, as the occurrence of one’s mental states is characteristically not of a piece with one’s awareness of them or indeed with being in a position to self-ascribe the relevant mental state. Indeed, to gain knowledge of them, we need to engage in observation and inference and we may fail to perform them, thus remaining blind to ourselves, as it were.

Still, there is an element of third-personal self-knowledge that sets it apart from knowledge of other people’s mental states. Namely, sometimes the inference can start on the basis of other mental states one has knowledge of in a first-personal way, such as one’s sensations, immediate feelings, propositional attitudes as commitments, etc.[[8]](#footnote-8) This obviously cannot be the case when we acquire knowledge of other people’s mental states through inference to the best explanation. In that case, the inference starts from observing their overt linguistic and non-linguistic behavior. Let us therefore consider some examples of third-personal self-knowledge.

Cases of inferences starting from the observation of one’s own behavior can be those that lead us to self-ascriptions of propositional attitudes as dispositions, whether of a Freudian kind or not. Consider a mother who reflects on her behavior towards her son and compares it with the one towards her daughter. She can judge that she is more often benevolent towards the latter and therefore infer that she prefers her daughter to her son. Of course, she may have gone wrong in judging that she is more munificent with her daughter in the first place, as she may under-estimate the number of times in which she has been open-handed towards her son. Hence, her self-ascription is neither groundless, nor transparent or authoritative.

Sometimes we figure out our complex emotions, such as love towards someone, by reflecting on our behavior and inner phenomenology in their presence or at the prospect of some event involving them. Obviously, we can go wrong in our self-interpretation. Our behavior and feelings may not be symptoms of love but of being possessive and thus dreading the prospect that a good friend should inevitably stop being so close to us as the result of getting married to another person. Alternatively, we can mistake our behavior and feelings as symptoms of love, when they are in fact signs of our vanity and complacent attitude towards those people who are usually kind to us (or to others in general). Sometimes, in contrast, we are self-blind in the sense that despite showing many of the characteristic symptoms of love, say, towards someone, we do not realize that we are in love with them. Such blindness can have different causes, both cultural and psychological.

Sometimes, we gain knowledge of our own minds, and in particular of our character traits, by means of testimony. That is to say, we form the belief “I am F”, where “F” is a psychological concept, because someone tells us we are *F*. This kind of third-personal self-knowledge, therefore, has an essentially social aspect to it and exceeds by far the case of the highly institutionalized communication that takes place between a patient and a psychological therapist. More specifically, it involves taking the other person at her word, in the absence of reasons to doubt of her competence and sincerity; or else, in light of positive reasons in favour of her competence and sincerity.[[9]](#footnote-9) A conversation with a friend, or a relative we have no reason to distrust or that we have positive reasons to trust and consider competent both in her mastery of the relevant psychological concept and in its application in our own case, can go a long way in giving us knowledge of our own minds. Still, we may be led astray and therefore err in our eventual psychological self-ascription.

Furthermore, there are cases in which we gain knowledge of our minds by means of inductive evidence. For instance, by knowing how we have reacted to certain situations in the past, we can inductively predict we will feel thus-and-so in the future, if the same conditions should obtain. Once more, this method may, on occasion, lead us astray, or we can fail to draw the relevant conclusions even if we have previous evidence which should justify us in forming the relevant belief. So, whenever self-knowledge is based on induction, groundlessness obviously fails, and so do transparency and authority.

We have also seen how alleged counter-examples to first-personal self-knowledge drawing on recent findings in cognitive studies can in fact be interpreted as interesting examples of third-personal knowledge of our own mental states. Therefore, for instance, we think we have certain character traits, or that we will have certain emotional responses in the light of alleged life-changing events, which we actually lack. The interesting aspect of this failure at self-knowledge is that the reason why we think we would have these mental dispositions is that we presumably engage in imagining how we would behave on a given situation. For whatever reason, which may or may not be biased, we think we would behave in ways that would depend on having a certain mental state. Yet, when faced with a real life situation, we do not behave as we imagined we would do. The kind of procedure followed in these cases, in order to try to gain knowledge of our mental dispositions, is indeed similar to the one we would apply if we were to figure out someone else’s mental states by means of simulation and its correct outcome is obviously not guaranteed. It is therefore another example of third-personal self-knowledge.

Another intermediate case, which does not immediately coincide with knowledge of our own mental states either through inference to the best explanation or through mental simulation, is the one in which we suddenly notice an aspect of ourselves we had been previously blind or oblivious to; or indeed reconceptualize an aspect of ourselves by suddenly seeing it in a different way. We may all of a sudden realize that we are not nice to other people—we do not actually care about them that much—but are only polite; or we may suddenly see a character trait of ours not as a symptom of strength but of weakness. Seeing-as, switches of aspect and noticing aspects are complex phenomena even when they clearly concern perceivable objects and properties and involve the operation of perceptual faculties.[[10]](#footnote-10) Obviously, in the case of our own mental states, these notions find only mediated or indirect employment, for no real perceptual faculty is involved and yet it does not seem that we engage in any complex inferential thinking, or in mental simulation. Furthermore, the relevant psychological concepts are necessarily needed, while, arguably,[[11]](#footnote-11) purely perceptual switches of aspects do not necessarily require concepts (although they may and often do involve them). We may say that we suddenly take a pattern of overt behavior, and possibly other inner aspects of our psychology, as instantiating this or that psychological property either because that pattern is manifest to us for the first time, or because the relevant concept is available to us at last and a pattern of behavior and inner elements we had possibly noticed before seems immediately to fit the newly acquired concept. Sometimes these new conceptualizations or reconceptualizations contradict previous ones and that gives rise to a switch of aspect regarding ourselves and our personalities. Still, in these cases, inference is involved at least in the following sense. While there is no inference to the best explanation, the characteristic notes of the newly acquired (or applied) concept guide the process of subsuming the observed pattern of behavior and inner elements under the concept. That may happen very rapidly, and so, phenomenologically, we may not even seem to be engaging in inferences, yet epistemically that is what grounds our judgment “I am F” (where F is a psychological concept). It is one possible case of self-interpretation and it can go wrong because, after all, the pattern observed may not fit the concept or we may have ignored other features, which would, if taken into account, lead to a different psychological self-ascription. The self-interpretation can be wrong for motivated reasons, often unconscious in nature, and this can explain at least some forms of self-deception.

What is important to notice is that third-personal self-knowledge exhibits all the typical aspects of any substantive cognitive achievement, which usually results in knowledge. We base our psychological self-ascriptions on various kinds of evidence, we engage in inferences and we deploy imagination to simulate a situation and predict our own actions, therefore determining their psychological motivations. All these procedures admit for the possibility of error and for the chance of not being deployed at all, thus resulting in a form of self-blindness. Furthermore, the kind of abilities called upon can be variously distributed among subjects, who are otherwise equal with respect to cognitive functions, conceptual endowment and rationality. Moreover, it can be finessed with experience, including exposure to literature, movies and drama, which very often affect us precisely because we identify with the characters and thereby get to know something about ourselves. Study too can augment our third-personal self-knowledge – especially when new concepts are acquired and make salient to us aspects of our mentality we were previously blind to. We can also improve over time by exercising these judgemental abilities and by learning to pay heed to relevant feelings and/or behavioural patterns and by becoming more and more proficient in the application of the relevant (dispositional) psychological concepts. Third-personal self-knowledge, therefore, is not any different – epistemically speaking – from knowledge of truths about physical objects around us, or about other people’s mental states in these respects. Yet, it is special in one kind of evidence it may derive from. That is, our first-personal knowledge of occurrent sensations, feelings, perceptual experiences and propositional attitudes as commitments. Yet, it is only for third-personal self-knowledge that the usual traits of knowledge are clearly present. For there is cognitive achievement, there is ample possibility of error, or self-blindness, without resulting in failures at rationality or ipso facto in lack of concepts’ possession, and there is room for expertise and for improving one’s skills. None of this, in contrast, seems to characterize first-personal self-knowledge. For some theorists, this means that first-personal self-knowledge is not, after all, real knowledge. Rather, it is a set of constitutive claims, which can be variously grounded and redeemed. I have explored the issue at length in Coliva (2016). For our purposes here, there is no need to take a stance on that.

What matters, rather, is to recognize that a powerful case can be made in favour of a pluralist conception of self-knowledge, which, as we have seen, comprises both first- and third-personal elements to it; and, in turn, for the view that third-personal self-knowledge is gained through a plurality of methods, which should be thoroughly investigated.

**References**

Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge (Mass.): Harvard University Press.

Burge, T. 2011. Self and self-understanding. The Dewey lectures (2007-2011). *The Journal of Philosophy* CVIII/6-7. Reprinted in *Cognition through Understanding*. Oxford: Oxford University Press: 140-228.

Cassam, Q. 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press.

Coliva, A. 2012. Human diagrammatic reasoning and seeing-as. *Synthese* 186/1: 121-148.

Coliva, A. 2015a. How to commit Moore’s paradox. *The Journal of Philosophy* 112/4: 169-192.

Coliva, A. 2015b. Review of Cassam’s *Self-Knowledge for Humans.* *Analysis* 1/1: 1-6.

Coliva, A. 2016. *The Varieties of Self-Knowledge.* London: Palgrave Macmillan.

Gertler, B. 2011. *Self-Knowledge.* London/New York: Routledge.

Gilbert, D. 2006. *Stumbling on Happiness.* New York: Knopf.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8: 529-566.

Nisbett, R. and T. Wilson 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84: 231-259.

Schwitzgebel, E. 2008. The unreliability of naïve introspection. *Philosophical Review* 117: 245-273.

Shoemaker, S. 1996. *The First Person Perspective and Other Essays.* Cambridge: Cambridge University Press.

Snowdon, P. 2012. How to think about phenomenal self-knowledge. In Coliva A. (ed.) *The Self and Self-Knowledge.* Oxford: Oxford University Press: 243-262.

Wegner, D. 2002. *The Illusion of Conscious Will.* Cambridge (Mass.): MIT Press.

Wegner, D. and T. Wheatley 1999. Apparent mental causation: sources of the experience of will. *American Psychologist* 54: 480-492.

Wilson, T. 2002. *Strangers to Ourselves. Discovering the Adaptive Unconscious.* Cambridge (Mass.): MIT Press.

Wittgenstein, L. 1953. *Philosophical Investigations.* Oxford: Blackwell.

Wright, C. 1998. Self-knowledge. The Wittgensteinian legacy. In Wright C., B. Smith and C. Macdonald (eds.) *Knowing Our Own Minds.* Oxford: Clarendon Press: 15-45.

1. Cf. Wright (1998), p. 16. [↑](#footnote-ref-1)
2. Similar considerations are advanced in Shoemaker (1996). [↑](#footnote-ref-2)
3. To be rational agents does not mean acting necessarily for good reasons. It means, however, to have knowledge of the actions one is performing and of their motivations, at least for the most part, so as to be in a position to be held responsible for them. [↑](#footnote-ref-3)
4. Whether, in turn, they are also constitutive elements of what it means to be a self or a subject at all, insofar as they are constitutive of being critical reasoners and subjects of moral norms as Burge (2011) maintains, is a further issue, which would need a separate treatment. [↑](#footnote-ref-4)
5. See Gertler (2011), pp. 70-86. [↑](#footnote-ref-5)
6. Gilbert (2006). [↑](#footnote-ref-6)
7. Nisbett and Wilson (1977); Libet (1985); Wegner (2002); Wilson (2002); Wegner and Wheatley (1999). [↑](#footnote-ref-7)
8. I was pleased to find a similar claim in Cassam (2014), reviewed in Coliva (2015b). [↑](#footnote-ref-8)
9. It depends on which theory of testimonial justification and knowledge one adopts, whether Humean in kind or Reidian. [↑](#footnote-ref-9)
10. Cf. Wittgenstein (1953, II, xi). [↑](#footnote-ref-10)
11. See Coliva (2012) for a discussion of seeing and seeing-as and the role of concepts in it. [↑](#footnote-ref-11)