Privileged Accessibility as Incorrigibility

Andrea Tortoreto

University of Turin

Abstract

This article investigates Katalyn Farkas's notion of privileged access as a criterion to distinguish the mental from the physical. Farkas argues that a state is mental if and only if its subject has a special kind of awareness of it, that is, if it has a unique subjective dimension. I compare this notion with Rorty's view that the mental can be characterized by incorrigibility, that is, being immune to third-person errors. I claim that the two notions are related but both have difficulties in accounting for the variety and intricacy of mental phenomena. In the final part of the paper, I not only analyze and contrast the views of Farkas and Rorty, but also suggest a modification of the concept of incorrigibility. In doing so, I attempt to provide a definition of the mental that is more adaptable and compatible with the variety and intricacy of mental phenomena.

Keywords: Privileged access, Incorrigibility, Mental states, Phenomenal character, Eliminative materialism.

1. Introduction

Katalyn Farkas' notion of privileged accessibility aroused great interest in recent debates in philosophy of mind (Farkas 2008). According to Farkas, the privileged accessibility that the subject has with respect to knowing itself is indeed the mark of the mental. Accordingly, for her a state is mental iff is accessed in a privileged way by its bearer, that is: a state is mental iff its phenomenal character includes an irreducible subjective component. She developes this account starting with a deep reinterpretation of cartesianism and a critique to Rorty.

Nevertheless, in the early Seventies, Rorty produced an interesting attempt to characterize the mental in terms of "incorrigibility", a notion really close to Farkas' idea of privileged accessibility. He (Rorty 1970, 1972) claims that first-person reports about thoughts and sensations are treated as incorrigible. He also argues that his specific concept of incorrigibility is the best candidate as the mark of the mental and the discriminating factor of the mental from the physical. On this basis he grounds his eliminative materialism, as he proceeds to suggest what sort of scientific development could eliminate the mental.

Argumenta (2024): 1—20
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.tor
First published: 10 December 2024

According to Rorty, in order to find the mark of the mental, it is necessary to begin by distinguishing two different notions of what counts as mental. The distinction Rorty has in mind is between mental events and mental features. In the first class fall thoughts and sensations considered as phenomenally aware occurrent states (). In the second class of mental states fall instead beliefs, emotions, intentions, desires, and so on.

On this groundwork, incorrigibility, that is the best candidate as the mark of the mental, can be applied only to the first kind of mental states, namely to mental events. Rorty definition of incorrigibility is as follows:

> A person R's belief that p at a time t is incorrigible iff there is no accepted procedure, from an empirical point of view, whose outcome would render rational to believe not-p at t.

My purpose here is twofold. In the first place, I shall argue that Rorty's incorrigibility is a particular kind of privileged accessibility and, like Farkas' notion, is totally different from the notion of infallibility. I am going to show that, with this in mind, Rorty's incorrigibility may still be promising in the contemporary debate about the mark of the mental. In this regard, I want to propose a refined definition of incorrigibility as a special form of privileged access, trying to find a synthesis between Farkas' and Rorty's positions: Assuming that each subject is deeply related to propositions that ascribe a phenomenally aware occurrent state to himself, it is then possible to consider a subject's belief that p at a time t incorrigible iff there is no accepted procedure, from an empirical point of view, whose outcome would render rational to believe not-p at t.

Secondly, I will though try to show that Rorty's analysis is off-target in eliminating the mental.

2. Privileged Accessibility

Katalin Farkas, in her well-known book titled The Subject's Point of View (Farkas 2008), suggests that Rorty is right in claiming that many of our intuitions about the mind are based on an uncritical reliance on the cartesian tradition. Yet, her aim is to defend and embrace this "elderly tradition", using Tyler Burge (Burge 1979: 73) definition, rather than overthrowing it. So, according to Farkas, "the conception of the mental we have inherited from Decartes may not be as easy to discard" as Rorty suggested and, on the contrary, it can be "fundamental to our understanding of ourselves as the kind of creatures we are" (Farkas 2008: 5).

To show this idea, Farkas starts off with a specific cartesian tool: the demon test. The word 'test' refers to Farkas' purpose to claim that there is a mark of the mental and to find it. To be sure, 'mark of the mental' is taken to mean the feature that makes it the case that all mental states have a unifying ground distinct from that of physical states. Anyway, to employ the demon test to find the mark of the mental does not automatically imply providing a reductive analysis of the notion of mind or even embracing ontological dualism. Farkas clearly states her will to remain neutral on the questions of dualism and physicalism and her aim to defend a conception of the mind perfectly compatible with both positions.

The cartesian account, as it is presented in the Meditations, eliminates nutrition and movement from the list of mental features, including on the contrary thoughts, emotions, and perceptual experiences. In Farkas' analysis of the cartesian argument, two candidates are considered as the mark of the mental and immediately refused. The first one is the idea of the "thinking thing" as a substance,

i.e. an immaterial mind independent of the physical. This concept is strictly connected with the idea that mental features can be exemplified even if nothing corporal exists. Farkas rejects this notion as it would contrast with at least some forms of physicalism.

The second proposal derives from an interpretation of the claim that the possession of mental states is believed with certainty, which means either with psychological indubitability or guaranteed correctness. The main problem with this notion, on both interpretations, is that it is possible to find cases where the possession of even non-mental features is believed with certainty, and where the possession of a mental feature is not believed with certainty. In other words, certainty cannot be considered the mark of the mental, because there are some judgements that the subject may access with a privileged epistemic stance but which do not involve mental states. Consider beliefs like: 'I am identical to myself', 'I exist', and so on—these are non-mental facts that I can attribute to myself with absolute certainty. On the other hand, I can be wrong with respect to my beliefs about my mental states, for instance in such cases like love (I can be wrong whether I love someone or not).

Once these options are discarded, Farkas clarifies her understanding of the role of the demon hypothesis. According to her, the role of the demon test is indeed "not to reduce the world to an incorporeal subject, but rather to reduce the world to a unique center of enquiry: to a subjective viewpoint" (Farkas 2008: 18). Viewed from this perspective, what "survives the introduction of the demon hypothesis is the subject and the portion of reality that is uniquely revealed from the subject's point of view" (Farkas 2008: 18). This subject's point of view is governed by a specific faculty, usually called 'reflection' or 'introspection', or, in Farkas' words, "special access".

Farkas' use of the demon device is therefore aimed at identifying the mental phenomena and their unifying ground. She finds the turning point in the notion of introspection. But what does Farkas means exactly with her term, 'special access'? It is important to note that she does not simply emphasize our faculty to know our mental features. To assume that the mental domain coincides with the introspectable domain is a substantive philosophical claim, but one that risks circularity. If we define mental states solely as those that are introspectable, and then assert that only introspectable states qualify as mental, this definition merely restates the initial assumption rather than offering an informative analysis of the mental domain. However, this assumption is not without controversy. Many philosophers argue that the mental domain should also encompass non-introspectable states, such as unconscious cognitive phenomena. The debate over whether introspection is a necessary or sufficient condition for mentality remains ongoing and unresolved in contemporary philosophy of mind. Farkas' claim, then, is that "introspection is the only cognitive faculty that provides privileged access to its subject matter" (Farkas 2008: 29). As I said, I am not completely sure that Farkas' argument is totally safe from the risk of circularity; what is absolutely sure though is that her treatment of subjectivity relies on a specific cognitive faculty and, for this reason, can be seen as an epistemic treatment.

Perhaps this point needs to be specified. One could in fact argue that Farkas' epistemic position about privileged access is ultimately grounded in her conviction that, phenomenologically speaking, mental states are states for a subject or, as Kriegel (2009) puts it, they are endowed with a subjective character. This may lead one to the suspicion that her position is only a pseudo-epistemical one.

My position is a little different. Farkas emphasizes that introspection provides privileged access to mental states, but this claim does not reduce her argument to a purely phenomenological position. According to Farkas, privileged access is a cognitive achievement grounded in the epistemic asymmetry between first-person and third-person access to mental states. It is this asymmetry that defines the epistemic nature of mental states, distinct from any phenomenological reduction of the mental to subjective experiences alone. While Farkas indeed recognizes that mental states possess a subjective component, this is incorporated into an epistemic framework that ensures that such states are not accessible to others in the same way they are to the subject.

Farkas herself refutes the idea that the mental realm is coextensive with the introspectable domain as a sheer tautology. Instead, her epistemic model proposes that the subject's point of view provides unique access to their mental states through introspection, and this access is a hallmark of mental states. This allows her to avoid a simplistic phenomenological reduction, while acknowledging that mental states, as perspectival facts, are necessarily accessed from the subject's unique vantage point.

By grounding privileged access in cognitive faculties, Farkas preserves the epistemic nature of mental states while also accounting for the subjective character that accompanies them. Therefore, the risk of Farkas' position being categorized as merely pseudo-epistemic is mitigated by her careful delineation of introspection as a form of knowledge acquisition that extends beyond subjective experience alone.

Farkas' account suggests that everything known through introspection belongs to the mind, and that privileged accessibility is the mark of the mental (Farkas 2010: X). In her view, the mental domain is characterized by a subject's unique access to her own mental states via introspection. This does not simply mean that mental states are introspectable, but that their introspectability signifies their belonging to the mental realm. My interpretation builds on Farkas' claim by emphasizing that privileged access, as described here, provides a foundation for distinguishing mental phenomena from other types of phenomena. So, the result of this account is that the unifying feature of the mental domain is the subject's point of view. Consequently, "to be a subject is to possess a point of view" (Farkas 2008: XIV) . Things appear to the subject in a certain way, under a unique light, and "this perspective includes not only the world around us, but also ourselves. There is a certain way for me to be when I am cold or when I am hot, when I am at ease or when I am worried" (Farkas 2008: 31). Therefore, to have a mind necessarily is to have such a point of view; while things simply surround objects, things are in a certain way for a subject.

Farkas' definition of the mental is grounded on two strong concepts. First of all, the idea that introspection is the only asymmetrical faculty, a faculty that allows us access to states or facts that no one else may enjoy. This grounds the claim that mental facts are always perspectival facts. The second point is that perspectival facts concern only phenomenal states in an extended sense of this term: in this sense, all conscious states have phenomenal properties. It follows that the phenomenal fully determines the mental.

In concluding this section, I will limit myself to mentioning an important possible objection to this account of the privileged access as the mark of the mental. This objection is that also proprioception provides for asymmetrical access to bodily states and that, therefore, I can not only have special access to my mental states, but I can also learn about my bodily states in a way no one else can.

If that is true, it is difficult to claim that privileged accessibility is the mark of the mental. Farkas' answer to this question is, again, a sort of cartesian answer. A certain bodily state, for instance a state of pain, gives rise to a specific feeling in my mind. This specific feeling is the subjective and particular way I learn something about my body: a way that is completely different from the way someone else would learn about that same state. The problem with this peculiar view is that it seems to force us into a dualist position: the subject has special access to the feeling caused by a bodily state, and this could be a way to radically separate the body and its states from the mind with its. In other words, trying to deny the asymmetricity of proprioception is not an easy task, and comes with the risk to deliver an ontological break between the mind and the body.

Probably Farkas could simply reply that, insofar as proprioceptive states are featured by privilege access, they are genuine mental states as well. But, while it is true that proprioceptive states may meet the criteria of privileged access, and therefore, under Farkas' framework, could be classified as genuine mental states, this leads to a broader question about the extension of the category of the mental. If proprioceptive states—traditionally viewed as physical sensations closely tied to bodily awareness—are to be included within the domain of the mental, this expands the category beyond what we might intuitively consider as mental phenomena, such as beliefs, desires, and emotions.

The potential inclusion of proprioceptive states forces us to reconsider whether the Mark of the Mental (MoM), defined purely in terms of privileged access, is sufficiently discriminatory. Although Farkas may consider proprioceptive states to be mental, this could potentially overly expand the definition of what constitutes the mental. It could encompass a range of states that do not share the same features commonly associated with paradigmatic mental states, like intentionality or subjectivity.

Therefore, the counterexample remains significant by revealing a conflict within Farkas' framework: either we broaden the mental category to encompass states like proprioception, challenging our conventional view of the mental, or Farkas needs to offer further criteria to differentiate between various types of states accessed through privileged means.

3. An Old Story

I argued that the notion of privileged access is vague in some way. This issue was at the center of a long debate in the analytical tradition, a debate that has mysteriously fallen into oblivion and was recently revived thanks to Farkas' book.[1] The point is: in which way one's access to one's own mental states might be thought as privileged? What type of epistemic superiority might be imputed to one's knowledge of one's own mental states?

In the literature, the first way of conceiving this favorable epistemic position is, probably, the cartesian idea that one's beliefs about his own mental states "cannot be false" (Descartes 1641: II.29). This idea is for instance taken up by Ayer, when he says that "I cannot be unsure whether I feel a headache, nor can I think that I feel a headache when I do not" (Ayer 1956: 55).[2] This is a way to conceive privilege

[1] For complete discussion about the first debate about privileged access see Alston 1971.
[2] A form of privileged access is also proposed by Locke: "a man cannot conceive himself capable of a greater certainty than to know that any idea in his mind is such as he perceives

access as a form of indubitability; a move, as I said, that Farkas explicitly refuses. But I think the problem must be investigated in a deeper way. It is possible to find, in the literature, at least two different senses of indubitability: a conceptual one and an epistemic one.

Malcolm is presumably the first to assert a conceptual form of indubitability. He says, for instance, that given the grammar of the word pain and the way in which we generally use such a word, it follows that "a) you can be in doubt as to whether I'm in pain, but I cannot; b) you can find out whether I'm in pain, but I cannot; and c) you can be mistaken as to whether I'm in pain, but I cannot" (Malcolm1967: 146).[3] According to Malcolm, because of the way we normally use the word 'pain', a statement like 'I doubt I'm in pain' is logically meaningless. That is, our use of the term 'pain' reflects a conceptual framework wherein the subject's report of pain is immune to doubt. This is not a matter of logical necessity in the strict sense, but rather an outcome of the way our linguistic practices structure the concept of pain.

In Malcolm's view, to express doubt about whether one is in pain would violate the very concept of pain as it is embedded in our linguistic framework. Thus, the indubitability of pain is a conceptual truth: it is built into the structure of how we understand and use the concept of pain within our language, rather than being a logical consequence deduced from abstract principles.

This conceptual form of indubitability is not acceptable to Farkas. The main reason to refute it is based on the observation that it is valid just for sensations but not for other mental states, e.g. love. But there is also a sense according to which indubitability can be conceived from an epistemic point of view.

An epistemic notion of indubitability is deeply grounded in the cartesian account. It is the idea of a rational impossibility of doubting any conscious state in its phenomenal nature. According to this view, it is not possible to entertain any doubt on a conscious state because a similar doubt calls into question consciousness itself. Here, impossibility is referred to the rational grounding of doubts, implying a sort of normative idea of impossibility;[4] to rationally admit the possibility of doubting phenomenal states is the same as to deny the existence of conscious states—which

it to be; and that two ideas, wherein he perceives a difference, are different and are not precisely the same" (Locke 1689: IV, 2).

[3] A very similar idea can be probably found also in Clarence Irving Lewis: "Subtract in what we say that we see, or hear, or otherwise learn from direct experience, all that conceivably could be mistaken; the remainder is the given content of the experience inducing this belief [...]. Apprehensions of the given which such expressive statements formulate, are not judgments, and they are not here classed as knowledge, because they are not subject to any possible error. Statement of such apprehension is, however, true or false; there could be no doubt about the presented content of experience as such at the time when it is given, but it would be possible to tell a lie about it" (Lewis 1946: 182-183).

[4] This way to frame impossibility is for example clearly stated by Sir Hamilton: "The facts of consciousness are to be considered in two points of view; either as evidencing their own ideal or phaenomenal existence, or as evidencing the objective existence of something else beyond them. A belief in the former is not identical with a belief in the latter. The one cannot, the other may possibly be refused [...]. Now the reality of this, as a subjective datum—as an ideal phaenomenon, it is absolutely impossible to doubt without doubting the existence of consciousness, for consciousness is itself this fact; and to doubt the existence of consciousness is absolutely impossible; for as such a doubt could not exist, except in and through consciousness, it would, consequently, annihilate itself" (Hamilton 1859: XV).

is in some way counterintuitive.[5] The point is that, contrary to the conceptual indubitability proposed by Malcolm, this version of indubitability as an epistemic privilege is something closer to Farkas' position. If I am so engaged with some propositions that whenever I believe one to be true there are no rational grounds to doubt that it is actually true, then I am really in a privileged epistemic position.

This idea of epistemic indubitability establishes a principle according to which a person is in a favorable position to discriminate true from false propositions concerning his present conscious states. And this is something that Malcolm's account does not allow. The fact that the logical structure of some words concerning phenomenally aware occurrent states is the ground for incorrigibility has nothing to do with the idea that a person is in an advantageous epistemic position about his conscious states.

From this point of view, I think Farkas is right in denying that privilege access could be considered a form of indubitability, but just if we look at indubitability in the conceptual sense. Epistemic indubitability is indeed perfectly coherent with her account. I believe that this is a crucial point, because this opens the door to Rorty's concept of incorrigibility, as I am going to consider in the next pages.

To sustain the indubitability thesis is indeed not the same to sustain the impossibility of mistakes. Rather, it is a weaker claim that there are no grounds for questioning the correctness of one's beliefs.[6] Before turning to Rorty's idea of incorrigibility, let us try to propose a first formulation of this concept that, in my opinion, could be perfectly in line with Farkas' notion of privilege access:

> (IT1). Each subject is so related to propositions that ascribe a phenomenally aware occurrent state to himself that it is impossible for him to believe that the same proposition is true and for someone else to prove that it is false.

The use of the term "proposition" is not casual: IT1 works independently from the natural language formulations uttered by the subject. This wording is consistent with Farkas' idea that to say that a person has privileged access to his phenomenally aware occurrent state is to say that his epistemic position regarding propositions ascribing phenomenally aware occurrent states to himself is advantageous in a way no one else's position is.

[5] Indeed, illusionist theories in the philosophy of mind, such as those advanced by Frankish (2016) and Kammerer (2019), argue that mental states are illusions generated by our cognitive architecture. These theories challenge the common-sense view that mental states such as thoughts, sensations, and emotions have a real, substantive existence. However, while I acknowledge the intellectual rigor behind the illusionist stance, I find it ultimately unconvincing. One of the key issues with illusionism is that it raises a significant challenge: if mental states are indeed illusory, how can we explain the deeply ingrained and seemingly inescapable experience of subjectivity? Even if mental states are reducible to neural processes or other physical mechanisms, their phenomenological presence—the what-it-is-like aspect of being a subject—remains a central feature of our experience. This is why I argue that, despite its appeal to some, the illusionist position faces significant obstacles in accounting for the richness of mental life: the persistence and significance of subjective experience seem to resist the idea that mental states are mere illusions.

[6] This idea first emerges in a passage from Ayer: "If this is correct, it provides us with a satisfactory model for the logic of the statements that a person may make about his present thoughts and feelings. He may not be infallible, but still his word is sovereign. The logic of these statements that a person makes about himself is such that if others were to contradict him we should not be entitled to say that they were right so long as he honestly maintained his stand against them" (Ayer 1963: 73).

To conclude this section, I want to briefly consider an important paper by Frank Jackson, in which he strongly defends the notion of incorrigibility against Armstrong's arguments. While I think Jackson's defense works, I don't think he is right in is definition of the incorrigibility thesis. According to Jackson, incorrigibility holds "that 'S believes at t that he is in pain at t' logically entails 'S is in pain at t'» (Jackson 1973: 51). Then, he defines privileged access as the different thesis according to which "a person's beliefs about his mental states may be false but cannot be shown false by anyone else" (Ibid.). I think there is great confusion at work here. First of all, Jackson's incorrigibility thesis is a very radical claim that has nothing to do with the way I considered incorrigibility in the previous passages. Jackson's idea fits rather with the concept of infallibility. The second point is that privileged access is not a different notion, but rather it is a general formulation so that incorrigibility is a kind of privileged access. This is naturally not a simple terminological problem but a theoretical one. Anyway, I think Jackson's arguments against Armstrong perfectly work also in defense of the true incorrigibility thesis, and that is something I will pick upon in the following sections of this paper.

4. Incorrigibility

As I said, Farkas' treatment of privileged accessibility derives from a deep critique of Rorty's philosophy of mind. Nevertheless, Rorty himself, in the early Seventies, produced an interesting attempt to characterize the mental in terms of incorrigibility. As mentioned, this notion is actually really close to Farkas' idea of privilege accessibility, as I will try to show. This attempt goes back to some deeply analytical papers produced by Rorty before the well-known book Philosophy and the Mirror of the Nature. In these papers, Rorty (1970, 1972 in particular) claims that first-person reports about thoughts and sensations are currently treated as incorrigible. He also argues that his specific concept of incorrigibility is the best candidate as the mark of the mental and the discriminating factor of the mental from the physical.

According to Rorty, in order to find the mark of the mental, it is necessary to begin by distinguishing two different notions of what counts as mental. The distinction Rorty has in mind is between mental events and mental features. In the first class fall thoughts and sensations considered as phenomenally aware occurrent states (). In the second class of mental states fall instead beliefs, emotions, intentions, desires, and so on. Rorty's idea is that "only the former class of mental entities generate the opposition between the mental and the physical, where this opposition is considered as an opposition between two incompatible types of entity, rather than an opposition between two ways of talking about human beings". The latter class of entities, indeed, are entities which "if we had never heard of thoughts and sensations, would never have generated the notion of a separate 'realm' at all". If, in fact, "we had no notion of a mental event, but merely the notion of men having beliefs and desires and, therefore, acting in such-and-such ways, we would not have had a mind-body problem at all" (Rorty 1970: 156). Thus, according to Rorty, believing and desiring, without any concept of mental events can be seen just as specific human activities. In this case, the only conceivable dualism could be between men as agents and men as mere bodies or, better put, between human beings who behave in a way explained by beliefs and desires and human beings who merely move their bodies. Thus, not a dualism between two different realms but between psychological or not-psychological ways of describing human behavior. If that's true, mental

events are, as a matter of fact, the "paradigm illustrations of what is meant by the Cartesian notion of the mental as a separate realm" (Rorty 1970: 156).

On this groundwork, incorrigibility, that is the best candidate as the mark of the mental, can be applied only to the firs kind of mental states, namely to mental events. Rorty definition of incorrigibility is as follows:

> (IT2). A person R's belief that p at a time t is incorrigible iff there is no accepted procedure, from an empirical point of view, whose outcome would render rational to believe not-p at t.

I call this definition IT2 and I believe it's a deepening of IT1, a more precise way to define incorrigibility as a form privileged access. A way which, it seems clear, is completely different from Jackson's definition. So, roughly speaking, a phenomenally aware occurrent state is mental iff at least one subject is in a position to have a belief about it whose epistemic authority cannot be empirically disproven.

Rorty assumes, contrary to Jackson, Quine's suspicion about the existence of necessities outside the natural realm and, in doing so, he proposes a definition of incorrigibility that does not refer to logical modalities. First of all, it is worth noting that this notion seems to be immune to Armstrong's classic arguments against incorrigibility. One well-known example is the so-called 'memory argument':

> Suppose I report 'I am in pain now'. If we take the view that the latter reports a piece of indubitable knowledge, to what period of time does the word 'now' refer? Not to the time before I started speaking, for there I am depending on memory, which can be challenged. Not to the time after I finish speaking, for then I depend on knowledge of the future, which can be challenged too. The time in question must therefore be the time during which the report is being made. But then it must be remembered that anything we say takes time to say. Suppose, then, that I am at the beginning of my report. My indubitable knowledge that I am in pain can surely embrace only the current instant: it cannot be logically indubitable that I will still be in pain by the time the sentence is finished. Suppose, again, that I am just finishing my sentence. Can I do better than remember what my stat was when I began my sentence? So to what period of time does the 'now' refer? (Armstrong 1968: 104-105)[7].

In this case, I fully agree with Jackson when he says that "the incorrigibility thesis is not a thesis about sentence tokens of certain kinds: it is not a thesis about particular utterances or inscriptions" (Jackson 1973: 53). As I said, incorrigibility is a thesis about propositions, unrelated to any particular natural language formulation. Thus, Armstrong's argument fails as propositions do not have temporal duration.

But there is also another important objection that must be considered in order to show why Rorty refuses a logical definition of incorrigibility. This objection comes again from Armstrong which assumes a Wittgensteinean[8] idea as follows:

> If introspective mistake is ruled out by logical necessity, then what sense can we attach to the notion of gaining knowledge by introspection? We can speak of gaining knowledge only in cases where it makes sense to speak of thinking wrongly that we have gained knowledge. In the words of the slogan: 'If you can't be wrong,

[7] Exactly the same argument is also presented in (Armstrong 1963: 420-421).

[8] This idea is imputed to Wittgenstein in Malcolm 1954 but it is possible to find a very similar argument also in Pitcher 1964 (280-285).

> you can't be right either'. If failure is logically impossible, then talk of success is meaningless (Armstrong 1963: 422).

This could be clearly a strong move against any logical treatment of incorrigibility, like Jackson's one, but surely not against Rorty's conception. In Rorty's account you can, at least in principle, be wrong about your mental states; what is not possible is an empirical proof against your thoughts and sensations. Thus, the first part of the claim is not consistent with the epistemic reading of incorrigibility as a kind of privileged accessibility: consequently, Armstrong's arguments do not work in principle.

Avoiding this kind of objections is, perhaps, one of the reasons that led Rorty to avoid purely logical treatments of incorrigibility. We can perfectly conceive that our consciousness can be deceived in some way, but this is not a problem if we consider incorrigibility just from an empirical and external point of view. The key point is that there is no safe way to go about fixing them, should they be in error. Viewed this way, this is to say that incorrigibility is considered to be a privileged status just from an epistemological perspective—as both Rorty and Farks would agree.

To sum up, epistemic incorrigibility circumvents the main objections to conceptual incorrigibility and, according to Rorty's account, "only mental events are the sorts of entities certain reports about which are incorrigible" (Rorty 1970: 166). This is not true of for a priori statements, simply because they are not reports, they are not descriptions of specific and particular states of affairs. Statements like '2+2=4' or 'Every event has a cause' are universal claims, not indeed descriptions of particulars.

But what about statements about appearances? When I say 'This looks brown to me now', it seems I am reporting a particular state of affair. Is this a report about a mental state? Rorty considers these kinds of statements quite ambiguous; they can be used to express a sort of hesitation: "To say that 'X looks brown' is, at the least, to express hesitation about saying that X is brown" (Ibid.). In this case surely 'X looks brown' is a report that someone is in a state of uncertainty about saying that X is brown. Yet it may not be, and it may simply be an expression of uncertainty and not a report. Now, for Rorty, only reports, or descriptions of particular states of affairs, are about mental states; statements of appearances can be reports, and therefore descriptions of the mental, but they can also be a mere abstention from judgement (in which case they are not reports).

At the end of these considerations, Rorty delivers a sufficient condition for something to be a mental event:

> If person R can have an incorrigible belief in some statement P which is a report on X, then X is a mental event.

It is clear that this condition are not applicable to mental features. The latter are such that "our subsequent behavior may provide sufficient evidence for overriding contemporaneous reports of them" (Rorty 1970: 167). Rorty's distinction between mental events and mental features is a useful framework for analyzing certain mental states, such as beliefs, desires, and intentions, which may be seen as implicit projections of future behavior. However, when it comes to emotions, the issue becomes more complex. Emotions, unlike beliefs or desires, have a distinct phenomenological quality, and one can feel an emotion like fear while acting contrary to it. This suggests that emotions may resist the same kind of predictive,

behavior-oriented analysis that works for other mental states. While Rorty's approach might place emotions within the realm of mental features, the unique phenomenological character of emotions warrants further investigation, as their connection to behavior may not be as straightforward. A more familiar distinction, as Farkas (2010) highlights, could be between conscious and standing states, where emotions might occupy a different status from beliefs and desires. Anyway, statements which could be considered as implicit projections of future behavior can be falsified by someone else. Such falsification provides an accepted procedure for overriding reports and, in doing so, provides also a distinction from reports of thoughts and sensations, which are independent from future behavior.

Thus, reports about mental features are corrigible, but, and that's a crucial point I need to stress, the chance of overruling reports about such features is realized only rarely. As Rorty suggests, "as such mental features as beliefs and desires become more particular and limited and, thus, approach the status of episodes rather than dispositions, they become more incorrigible" (Rorty 1970: 168).

In this sense, there's no strong distinction between saying that I am afraid of the green slime I just met and saying I had a sense of fright when I met the green slime. Momentary beliefs, desires, and emotions tend to fall into thoughts and sensations, thus becoming more like episodes than dispositions. From this point of view, mental features are almost incorrigible: they tend to become incorrigible as they become more particular and restricted. Rorty thus speaks of "near-incorrigibility" for reports of mental features and "strict-incorrigibility" for reports of mental events. The conclusion is that the latter is the mark of mental events; near-incorrigibility, by having a family resemblance with strict-incorrigibility, proves that incorrigibility as a general notion ties together everything we consider mental things.

On this basis, we may derive that Rorty's treatment of the mark of the mental comes to a weaker notion of what makes a state or a property count as mental. A notion that, however, is still promising in today's debate because this weakness makes it quite flexible. If, on the one hand, it is true that Rorty's epistemic incorrigibility fails to provide a definite set of necessary and sufficient conditions for mentality,[9] on the other hand, family resemblance between strict- and near-incorrigibility still grounds a unifying factor for everything mental against the physical.

To sum up the discussion presented so far, I want to refine the concept of incorrigibility as a special form of privileged access, trying to find a synthesis between Farkas' and Rorty's positions:

[9] Rorty's proposal is ultimately framed as a conditional rather than a biconditional. That is, he specifies conditions under which certain events *could* be classified as mental, but he does not really claim, contrary to what he intended, that these conditions are *both* necessary and sufficient for all mental phenomena. Rorty's focus is primarily on what he terms "mental events", which form a limited subset of the broader category of mental states. These are phenomena that display certain features like incorrigibility and epistemic access from the first-person perspective. However, this framework leaves out a significant portion of what we generally consider mental states, including long-standing dispositions, attitudes, and even emotions that do not always meet the same criteria. This limitation in Rorty's account is key to understanding why his delivery fails to provide a full set of necessary and sufficient conditions. By confining his analysis primarily to mental events, Rorty does not account for mental states that lack the episodic nature of such events, including standing beliefs, desires, and other mental features that exist even when we are not consciously aware of them. As a result, his account offers at best a partial analysis of the mental domain, and this is why I argue that his conditions apply only to a subset of mental phenomena, rather than providing a comprehensive definition of mentality.

(IT3). Assuming that each subject is deeply related to propositions that ascribe a phenomenally aware occurrent state to himself, it is then possible to consider a subject's belief that p at a time t incorrigible iff there is no accepted procedure, from an empirical point of view, whose outcome would render rational to believe not-p at t.

5. Graduate Incorrigibility

Rorty's approach to the question about the mark of the mental is indeed, like Farkas', an epistemological approach. The existence of the mental itself derives from the fact that our linguistic community accords a peculiar, epistemic privilege to some mentalistic propositions. From this point of view, Rorty's treatment, as I said, is even more precise, as it avoids a certain vagueness attached to the notion of privileged accessibility; moreover, he successfully puts aside the risks of falling back in dualistic positions. The notion of incorrigibility, I tried to specify with IT3, is a possible way to develop privileged access by showing how it works pragmatically. From this point of view, Rorty's take on incorrigibility seems to be a still promising account for the contemporary debate.

So, in the current debate on the Mark of the Mental, I argue that incorrigibility offers a more robust and precise criterion for distinguishing mental states than privileged access. One of the core motivations for this claim lies in the observation that privileged access, as traditionally conceived, suffers from an inherent vagueness when applied to certain mental phenomena. Privileged access suggests that we can know our mental states directly and non-inferentially, but this notion remains susceptible to varying interpretations, particularly regarding the degree of awareness and the nature of the introspective process.

For instance, Farkas defines privileged access in terms of a subject's ability to have immediate, first-person knowledge of their mental states. While this idea captures the intuitive sense that mental states are intimately known by the subject, it lacks precision when we consider cases of mental states that may be known indirectly or through reflection rather than immediate introspection.

In contrast, incorrigibility provides a clearer and more defensible criterion by focusing on the impossibility of error in the subject's report of their mental states. The key advantage of incorrigibility is its stability: even if external factors allow another person to access or perceive the same mental state, the subject's report of their experience remains privileged in the sense that it cannot be overridden by external corrections. This ensures that the mental state retains its subjective, first-person character, which privileged access alone does not necessarily safeguard.

The distinction between privileged access and incorrigibility becomes especially significant when addressing counterexamples like the one involving the Hogan twins or futuristic machines capable of linking experiences. The case of the Hogan craniopagus twins, who share a thalamic bridge and report direct awareness of each other's sensations, provides a challenging counterexample to Farkas' notion of privileged access as the MoM. This counterexample could be also translated in science fictional terms; let's consider for instance a futuristic machine able to establish similarly immediate connections between two subjects' experiences. According to Farkas, introspection grants individuals a form of access to their mental states that no one else may enjoy (Farkas 2010: 4). However, the Hogan twins, who directly share the same sensations, seem to challenge this assumption. If two subjects can access the same mental state, can we still regard that state as

exclusively privileged? Or, could a certain mental state M stop being mental just because someone else acquires the same kind of privileged access to it?

The strength of incorrigibility, when compared to privileged access, lies in its ability to accommodate such scenarios without compromising the mental status of the state in question. Even though both twins may share the same direct access to a given mental state, there would still be no empirical procedure capable of rationally correcting either of them about that state. In this sense, the state remains incorrigible for both subjects, even though it is no longer unique to one individual; there will still be no empirical procedure for making it rational to believe that not-M is experienced since the twin/machine-hooked-person will entertain the very same M.

Thus, the Hogan twins scenario suggests that incorrigibility offers a more stable definition of mentality than privileged access. Regardless of whether another subject shares access to the same mental state, the incorrigibility of the mental state remains intact: no external observation or empirical method can invalidate the subject's report of that state. This supports the claim that incorrigibility not only clarifies the notion of privileged access but also provides a more robust and adaptable framework for defining what it means to have a mental state.

So, concluding this comparison, it is possible to say that while both Rorty's incorrigibility and Farkas' privileged access offer plausible frameworks for understanding the mark of the mental, both positions encounter difficulties when applied to the full range of mental phenomena. Farkas' view rests on the idea that introspection provides privileged access to one's mental states. This conception, however, faces issues when considering phenomena that extend beyond simple introspection. For instance, proprioceptive states—those related to the body's awareness of itself—do not fit neatly into the introspective model, yet they are often considered part of the mental. If proprioception is characterized by privileged access, then it should, according to Farkas' framework, count as genuinely mental. However, this expands the domain of the mental in ways that are not fully accounted for in Farkas' epistemic model.

Rorty's criterion of incorrigibility, while providing a more precise framework, also struggles with the diversity of mental phenomena. As noted, incorrigibility offers a robust definition for mental states by positing that the subject's report of their mental state cannot be rationally corrected by empirical observation. However, this view does not fully explain how we might deal with mental phenomena that resist such epistemic certainty, such as unconscious cognitive states (e.g., blindsight) or the complexity of emotional experiences, which may not lend themselves to clear, incorrigible reports.

Moreover, both approaches have difficulties in accounting for the intricacy of shared introspective experiences. As pointed out, the case of the Hogan craniopagus twins challenges the notion of exclusive introspective access. If two individuals can share direct access to one another's sensory experiences, does this mean that mental states cease to be mental simply because they are shared? Incorrigibility offers a clearer resolution to this problem, since the mental state remains incorrigible even when shared between two individuals. However, this leads us to reconsider the nature of privileged access itself and the need for a more flexible, comprehensive understanding of the mental that can accommodate such complexities.

In this context, I think my attempt to refine the notion of incorrigibility leads to a more adaptable and nuanced definition of the mental. This refined definition is not only epistemically stable but also flexible enough to account for the full

range of mental phenomena, from proprioception to shared experiences, without relying solely on introspective exclusivity.

Allow me to elaborate on this point for a moment, with some suggestions for possible future analyses. One of the key challenges in defining incorrigibility as the MoM, as I said, lies in accommodating the vast diversity and complexity of mental states. While traditional conceptions of incorrigibility have focused on states like pain or simple perceptual experiences, where a subject's report is taken as necessarily true, many mental states are far more complex and resistant to this simplistic model. To account for the variety of mental phenomena, it is probably necessary to propose a more flexible and graduated notion of incorrigibility—one that can handle both simple and more intricate mental experiences without losing its core epistemic insight.

First, let us consider simple states, such as physical pain or basic emotions like fear or anger. These states have typically been regarded as paradigmatic cases of incorrigibility: when a subject reports feeling pain, it seems absurd to deny that they are indeed experiencing that pain. These states are easily categorized as incorrigible because they involve a direct, first-person access that cannot be overridden by third-party observation. The certainty with which one experiences such states provides the foundation for the traditional understanding of privileged access and incorrigibility.

However, this model begins to falter when we consider more complex mental states, such as mixed emotions, moods, or semi-conscious experiences. For example, one might feel a vague sense of melancholy without being able to precisely identify the cause or even the full nature of that emotion. In such cases, the subject's access to the mental state is still privileged, but it is less clear-cut than in the case of sharp, acute pain or fear. Here, a more nuanced understanding of incorrigibility is needed—what we might call qualified incorrigibility—where the subject has privileged access to the existence of the mental state, but not necessarily full or accurate access to its nature or causes. This gradation allows for a more flexible application of the concept across different types of mental experiences.

Furthermore, certain mental states, such as proprioceptive awareness (the awareness of one's body's position and movement), challenge the traditional notion of incorrigibility. While these states are only accessible to the subject, they are often not as clear or definitive as pain or basic emotions. For example, one may be aware of a slight tension in the body but unsure whether it is discomfort, stress, or fatigue. In such cases, we can again invoke the notion of partial incorrigibility—the subject has privileged access to the fact that he is experiencing a sensation, but not necessarily to its full nature or meaning. This expansion allows us to maintain incorrigibility as the MoM while recognizing the complexity and ambiguity inherent in certain types of mental states.

This flexible interpretation also extends to unconscious states, such as blindsight, where the subject reacts to visual stimuli without conscious awareness. While the subject lacks direct, conscious access to the stimuli, the information still plays a role in his mental life, influencing behavior and decision-making. For instance, a person with blindsight is able to react to visual stimuli in their blind field, even though they do not have conscious visual awareness of those stimuli.

A more graduated concept of incorrigibility could account for these cases by recognizing that even unconscious or semi-conscious mental states exhibit a form of privileged access that is not fully accessible to others, though it may not rise to the level of traditional conscious awareness; blindsight and similar phenomena

are indeed mental states despite not being introspectively accessible in the way more traditional mental states (e.g., emotions or beliefs) are.

In this sense, I propose that such states exhibit a form of quasi-incorrigibility: although they are not introspectively incorrigible in the traditional sense, they still resist correction by empirical means, as the subject is unaware of the internal processes that guide their behavior in response to these stimuli. The fact that blindsight patients are able to make accurate judgments about visual stimuli they claim not to see demonstrates that these unconscious states maintain a certain epistemic stability—they are, in some sense, incorrigible in their capacity to influence behavior without being accessible to conscious reflection.

In light of these considerations, I argue that the traditional conception of incorrigibility as a strict, all-or-nothing feature of mental states is too narrow to account for the full range of mental phenomena. By introducing the ideas of qualified incorrigibility, partial incorrigibility, quasi-incorrigibility or graduate incorrigibility, we can develop a more flexible and robust understanding of incorrigibility that preserves its central role in defining the mental, while accommodating the complexity and diversity of mental experiences. This revised framework enables us to maintain the epistemic authority of the first-person perspective, even in cases where that perspective is ambiguous, shared, or partially unconscious. Ultimately, this broader conception of incorrigibility offers a more stable and adaptable definition of mentality, one that can withstand challenges posed by both contemporary cognitive science and philosophical thought experiments.

The definition I proposed in IT3 is centered on the idea that incorrigibility is tied to the absence of any empirical procedure that could make it rational to disbelieve the subject's belief about their own mental state. In other words, incorrigibility depends on the impossibility of empirical correction.

The gradual notion of incorrigibility I'm imagining as a possible development introduces more flexibility in terms of the clarity and exclusivity of access to mental states, but it maintains the core epistemic feature emphasized in IT3: namely, that there is no external, empirical procedure that could override the subject's first-person access to their mental state. The main difference lies in the type of access and the certainty involved in different mental states, rather than in undermining the central epistemic criterion of incorrigibility itself.

Thus, this expanded version seems to be compatible with my definition, as both maintain that incorrigibility is grounded in the impossibility of empirical disconfirmation. What the expansion adds is a more nuanced understanding of the types of mental states and the degrees of certainty that different kinds of mental experiences may involve. This does not conflict with IT3, but rather complements it by showing how even complex, ambiguous, or shared mental states can still be protected from empirical correction in the way IT3 describes.

6. Folk Psychology

Anyway, in Rorty's project there is a very problematic point. It is well known that, in these early analytical papers, Rorty's aim is to develop a form of eliminative materialism that will have a great impact in the following years. On the grounds of the concept of incorrigibility, he further suggests what sort of scientific development would be sufficient to outright eliminate the mental. I think this inference is not justified.

Rorty's idea in these early papers is that future developments in the cognitive sciences and neurosciences will enable us to eliminate mentalistic language and, in doing so, first-person reports about thoughts and sensations which are currently treated as incorrigible. In this way, the disappearance of the mark of the mental will coincide with the elimination of the mental tout court. I will thus try to argue that Rorty's conception of the relationship between neurosciences and eliminative materialism is incorrect.

Now, according to Rorty's argument, the nature of the mental must be characterized in terms of incorrigibility. However, there are some scientifically sufficient conditions for the elimination of such incorrigible statements and, consequently, of mentalistic language (a precondition for defending materialism). Once these conditions have been met, physicalistic objects and materialistic language will carry out the task, hitherto realized by folk psychology, in a better and more adequate way. Mental reports will be no more incorrigible when mental states will be definitively explained by observation of the physical inner workings of the brain. I believe this conclusion is too simplistic and not justified by a more complete consideration of how our language practices really work. Suppose, for instance, that neurological investigation informs us that a subject has a thought of her mother, even though the same subject reports that she is 'not thinking of anything. Even in this case, there still remain some facts that ought to be explained through mentalistic language: what we might want to understand is why the subject in question is thinking of her mother and, specifically, why she has lied about it. In other words, even after thorough scientific investigation, a set of questions leading us back to the mentalistic vocabulary is still inescapable.[10]

But it is also possible another compelling counterexample from an opposite point of view, following Wittgenstein's Zettel. Wittgenstein discusses how a person might claim to be in pain or have a subjective experience, without the terms relating to sensations, used by that same person, referring to something internal. Today, in a wittgesteinian sense, we could say that someone might claim to be in pain even when there is no observable physical evidence of such a state in the brain. In a modern neuroscientific context, imagine a scenario where a patient consistently reports intense pain, but advanced brain imaging techniques fail to reveal any correlating neural activity in the regions typically associated with pain processing.

This scenario raises a fundamental question: Is the patient not in pain simply because neuroscience fails to detect it? According to Rorty's materialist perspective, where mental states are presumed to be fully reducible to brain states, this absence of physical evidence should lead us to doubt or even deny the reality of the reported pain. However, Wittgenstein's analysis suggests otherwise. Wittgenstein famously says, "It cannot be said of me at all (except perhaps as a joke) that I know I am in pain. What is it supposed to mean—except perhaps that I am in pain?" (1953: §246). The point here is that first-person experience is irreducibly authoritative for the subject who experiences it, even in the absence of third-person verification.

When I express my sensation of pain—whether by stating 'I am in pain' or through groaning—I am not sharing my knowledge of the sensation, but actually revealing the pain itself and seeking acknowledgment from others. When someone reacts to my "pain behavior", for example, by comforting me with words like

[10] A very similar version of this argument was proposed by Doppelt (1977: 532-534).

'I know you're in pain', they aren't merely conveying their understanding of my private sensation. Instead, they are showing empathy, recognizing that it is I who am in pain. In doing so, they affirm the significance of our relationship and the importance of acknowledging my suffering (Wittgenstein 1967: §482-435).

This example demonstrates that the subjective, first-personal access to mental states, like pain, cannot be fully captured by external, third-person scientific investigation. It echoes the idea that the phenomenological dimension of mental states resists straightforward reduction to physical processes. This resistance calls into question the eliminative materialist thesis, which assumes that mental phenomena can be fully explained away by neuroscientific data. Wittgenstein's thought experiment illustrates that a person's subjective experience holds epistemic primacy over external, observational data, meaning that the mental cannot be dismissed simply because the physical is absent or undetectable.

Moreover, this leads us back to the concept of incorrigibility, which was central to Rorty's epistemic criterion. Incorrigibility refers to the idea that certain mental states are beyond correction from external observation. For example, if someone sincerely claims to be in pain, no external observer can definitively refute this claim, regardless of what neurological evidence (or lack thereof) they may present. Rorty himself endorsed this criterion but took it in a direction that eventually led to his eliminativist stance—arguing that the incorrigibility of mental states might render them illusory and subject to elimination as scientific knowledge advances. However, the Wittgensteinian counterexample demonstrates that this incorrigibility speaks not to the non-existence of mental states, but rather to their fundamentally distinct epistemological status from physical states.

The example of a patient feeling pain despite a lack of neurological evidence underscores the robustness of first-personal, subjective reports in defining the reality of mental states. Pain, as an experiential reality, exists for the person who experiences it, regardless of whether external instruments can measure it. Thus, this counterexample shows that there is something irreducible about the mental that remains resistant to physicalist explanations.

By introducing this example, it becomes clearer that Rorty's form of materialism is incomplete. It fails to account for the nuanced and sometimes unobservable nature of mental states as they are experienced from the first-person perspective. In this way, Wittgenstein's critique highlights the limitations of materialism in explaining the full range of mental phenomena, particularly in cases where subjectivity plays a central role.

It ultimately seems that Rorty's picture of scientific progress is too simple and does not take into account the impossibility of a complete elimination of teleological paradigms to explain human actions. Our scientific framework, in other words, appears to be inextricably connected to mentalistic vocabulary. We cannot over-simplify its relationship with the mental as an absolute clash. A clear contingent proof of this is that incorrigibility and privileged accessibility are still central topics in contemporary philosophy of mind.

A perfect discovery of mental-physical correlations, in Rorty's sense, does not seem sufficient to delate our commitment to the mentalistic language. To share a better understanding of this point I think it is important, from a theoretical point of view, to go back to Sellars. Sellars' work had a great influence on the early Rorty's papers,[11] and some central remarks in Empiricism and the Philosophy of Mind must

[11] For this point see Auxier and Hahn 2010: 8-10.

be considered as a conceptual ground for eliminative materialism and, in particular, for the purposes of this paper. Sellars dose not embrace eliminative materialism per se; however, he does provide a theoretical anchor to Rorty's account in those passages of his masterpiece where he shows the possibility to frame folk psychology as a theory.

It is not by chance that recent proponents of the so-called 'theory theory of mind' refer to Sellars as a fundamental source. In the famous thought experiment known as the Myth of Jones, Sellars tries to show that thoughts are postulated entities theorized starting from the language used to explain manifest behaviors. Sellars holds that behaviorism is a useful methodological device that delivers an analysis of the concepts of mental states. Sellars is anyway committed to the existence of internal states, but he also believes it is possible to think of mentalistic concepts in an analogous way, with respect to the concepts pertaining to verbal behavior. From this perspective, verbal behaviorism can be used to examine and reconstruct concepts relating to intentional states.

On the one hand, Sellars rejects the idea that it is possible to exclude all reference to private and internal states from language. On the other hand, however, he believes it is possible to explain the origin and the nature of internal states through the idea that they are dependent on the most primitive intersubjective discourse, hence on manifest behaviors.[12] Sellars describes a prehistoric society in which men do not have a mentalistic vocabulary. That is, its members can only speak of observable behavior without any reference to internal states: they do not possess terms that refer to desires, volitions, intentions, and so on. At some point in the story, a brilliant man called Jones enters the scene. He notes that people behave in a rational and intelligent way even if they are not talking. Sometimes they perfectly describe what they are doing (they think out loud, in sellarsean language) and sometimes they just act without saying anything. Nevertheless, their actions are guided by coherent and perfectly understandable motivations. When they explain what they are doing, their fellow citizens can perfectly and completely understand their reasons, but if they remain silent, their actions seem to be inexplicable; this happens because Sellars' prehistoric community lacks mentalistic language. To solve this inconsistency, Jones formulates a genial theory: he postulates the existence of internal episodes that have the same structure as spoken language. Jones calls these episodes thoughts. Thus, overt utterances are the final step of a process which begins with internal episodes and his model for these kinds of episodes is the same overt verbal behavior. So, Jones' theory is that overt verbal behavior is the manifestation of thoughts and, on the ground of this theory, Jones himself trains his fellow cavemen to understand and interpret others' behavior. Yet, this is not the end of the story, because, as Sellars points out, it takes just a short step to use the same language for self-description. Over time, the ability to do so spreads, and quickly the members of prehistoric society becomes more and more expert at describing their own mental states. In this way, "what began as a language with a purely theoretical use has gained a reporting role" (Sellars 1956: 188).

The Myth of Jones is a handy story to illustrate that folk psychology, the standard commonsense framework through which we talk about desires, beliefs, dreams, emotions, and so on, is not a brute given, but rather a theory we apply

[12] This is the main topic of the famous correspondence with Chisholm. See Chisholm-Sellars 1972.

over certain data. This way of looking at folk psychology, and that is the point I would like to stress, has a crucial influence on Rorty. If we think that folk psychology, and all its linguistic baggage, is just a theory, we can also believe that it is falsifiable like all scientific theories, at least in principle. This idea, I think, is the ground to make the move that Sellars does not make; this move is called eliminative materialism. If folk psychology is a falsifiable theory, the entities it postulates, namely internal episodes or thoughts, might not exist.

Anyway, the step Rorty has taken is far too long. Rorty's belief that our linguistic community accords a special epistemic status to mentalistic utterances is indeed not enough to eliminate thoughts: it is rather a way of arguing, as Sellars shows, that our theories about thoughts are based on the natural language we use to describe shared behaviors. The core of the Myth of Jones is not that folk psychological language is eliminable, but rather that "concepts pertaining to such inner episodes as thoughts are primarily and essentially intersubjective" and that the reporting role of those concepts, based on the fact that each of us has a privileged access to his own thoughts, "constitutes a dimension of the use of these concepts which is built on and presupposes this intersubjective status" (Sellars 1958: 188).

In this sense, the intersubjectivity of language frameworks is perfectly compatible with the privacy of internal episodes. Rorty himself, in fact, eventually understood this in his later works; however, his misinterpretation in the Seventies was led by linking the notion of incorrigibility, an interesting way to characterize privileged accessibility, to eliminative materialism.

The Myth of Jones, connected with Rorty's treatment of incorrigibility, shows that the existence of the mental is grounded on the fact that our linguistic community accords a special epistemic status to mentalistic utterances. But, as I said, this does not mean that thoughts and sensations can be eliminated. It just means that mental states, and their incorrigibility, are linguistically accessible and, consequently, that our beliefs are socially construed. Something really connected with the inferentialist view proposed by some sellarsian scholars, but also with the neo-pragmatism suggested by the later Rorty.

References

Alston, W., 1971. Varieties of Privileged Access. *American Philosophical Quarterly*, 8, 223-241.

Armstrong, D.M., 1963. Is Introspective Knowledge Incorrigible?. *Philosophical Review*, LXII, 417-432.

Armstrong, D.M., 1968. *A Materialist Theory of Mind*. London: Routledge.

Auxier, R. and Hahn, L., 2010. *The Philosophy of Richard Rorty*. Chicago: Open Court.

Ayer, A.J., 1956. *The Problem of Knowledge*. London: Macmillan & Co..

Ayer, A.J., 1963. Privacy, in Id., *The Concept of a Person and Other Essays*. London: Macmillan & Co..

Burge, T., 1979. Individualism and the Mental. *Midwest Studies in Philosophy*, 4(1), 73-122.

Descartes, R., 1641/2008. *Meditations on First Philosophy*. New York: Oxford University Press.

Doppelt, G., 1977. Incorrigibility, the Mental and Materialism. *Philosophy Research Archives*, 504-536.

Farkas, K., 2008. *The Subject's Point of View*. Oxford: Oxford University Press.

Framkish, K., 2016. Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, 23, 11-39.

Hamilton, W., 1859. *Lectures on Metaphysics*. Vol. I, Edinburgh: William Blackwood and Sons.

Jackson, F., 1973. Is there a Good Argument Against the Incorrigibility Thesis?. *Australian Journal of Philosophy*, 51, 1, 51-62.

Kammerer, F., 2019. The Illusion of Conscious Experience. *Synthese*, 198(1), 845-866.

Kriegel, U., 2009. *Subjective Consciousness. A Self-Representational Theory*. Oxford: Oxford University Press.

Lewis, C.I., 1946. *An Analysis of Knowledge and Valuation*. Le Salle: Open Court.

Locke, J., 1689/1998. *An Essay Concerning Human Understending*. London: Penguin.

Malcolm, N., 1954. Wittgenstein's Philosophical Investigations. *The Philosophical Review*, 63, 4, 530-559.

Malcolm, N., 1967. The Privacy of Experience. In: A. Stroll, ed. *Epistemology: New Essays in the Theory of Knowledge*. New York: Harper and Row, 129-158.

Pitcher, G., 1964. *The Philosophy of Wittgenstein*. Englewood Cliffs: Prentice-Hall.

Rorty, R., 1970. Incorrigibility as the Mark of the Mental. *Journal of Philosophy*, LXVII, 12, 399-424, reprinted in R. Rorty, Mind, *Language and Metaphilosophy. Early philosophical papers*. Cambridge: Cambridge University Press, 2014.

Rorty, R., 1972. Functionalism, Machines, and Incorrigibility. *Journal of Philosophy*, LXIX, 8, 203-220, reprinted in R. Rorty, *Mind, Language and Metaphilosophy. Early philosophical papers*. cit.

Rorty, R., 1981. Is there a Problem about Fictional Discourse?. In: D. Heinrich, W. Iser, eds. *Funktionen des Fictiven: Poetik und Hermeneutik*, Munich: Fink Verlag, reprinted in R. Rorty, *Consequences of Pragmatism*. Minneapolis: University of Minnesota Press, 1982, 110-138.

Sellars, W., 1958/1991. Empiricism and the Philosophy of Mind. *Minnesota Studies in the Philosophy of Science*, 1, Minneapolis: University of Minnesota Press, reprinted in W. Sellars, *Science, Perception and Reality*. Atascadero (CA): Ridgeview Publishing Company.

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

Wittgenstein, L., 1967. *Zettel*. Oxford: Basil Blackwell.