Theoria, 2023, 38(1), 5-24
https://doi.org/10.1387/theoria.23828

Ὁ Θεὸς ἀριθμητίζει

THEORIA

ISSN 0495-4548 – eISSN 2171-679X

Mental expression and inner speech

(La expresión mental y el habla interna)

Jesús López Campillo*
Universidad de Murcia

ABSTRACT: This article explores the importance of mental expression in understanding the phenomenon of inner speech. Most accounts of inner speech assume from the outset the common idea that the expressions of a subject (e.g., a smile) and their mental states (e.g., joy) are two different types of items somehow related to each other. This relational view of expression is challenged in this article. Firstly, it is argued that relational views of expression cannot explain some features of inner speech. Secondly, a non-relational view of expression is developed, according to which mental states are patterns of expressive behavior. Thirdly, it is argued that only from the framework of non-relational expressivism is it possible to explain the main features of inner speech. Finally, it is concluded that non-relational expressivism emerges as a prominent contender among contemporary views of the mind, as it provides the only account of inner speech that can fully explain the phenomenon.

KEYWORDS: Mental expression, inner speech, mental states, expressive behavior, non-relational expressivism, mind.

RESUMEN: *Este artículo explora la importancia de la expresión mental para comprender el fenómeno del habla interna. La mayoría de las concepciones del habla interna parten de la extendida idea de que las expresiones de un sujeto (ej., una sonrisa) y sus estados mentales (ej., alegría) son dos tipos de ítems distintos relacionados entre sí de alguna manera. Esta visión relacional de la expresión es puesta en cuestión en este artículo. En primer lugar, se argumenta que las concepciones relacionales de la expresión no pueden explicar algunas características del habla interna. En segundo lugar, se desarrolla una concepción no relacional de la expresión, según la cual los estados mentales son patrones de comportamiento expresivo. En tercer lugar, se argumenta que las principales características del habla interna sólo pueden ser explicadas desde la perspectiva de una concepción no relacional de la expresión. Por último, se concluirá que el expresivismo no relacional emerge como un contendiente sólido entre las diferentes concepciones contemporáneas de la mente, ya que es la única concepción que puede ofrecer una explicación completa del fenómeno del habla interna.*

PALABRAS CLAVE: Expresión mental, habla interna, estados mentales, comportamiento expresivo, expresivismo no relacional, mente.

**** Correspondence to:*** Jesús López Campillo. Philosophy Department, Faculty of Philosophy, University of Murcia, Avda. Teniente Flomesta, N.º 5, Ed. Convalecencia, 30003, Murcia (Spain) – jesuslcampillo@gmail.com – https://orcid.org/0000-0003-0216-5637

How to cite: López Campillo, Jesús. (2023). «Mental expression and inner speech»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 38(1), 5-24. (https://doi.org/10.1387/theoria.23828).

Received: 19 July, 2022; Final version: 20 March, 2023.

ISSN 0495-4548 - eISSN 2171-679X / © 2023 UPV/EHU

1. Introduction

It is almost a truism that mental states are expressed in linguistic and non-linguistic behavior. Asserting "Today is Sunday" can be an expression of belief, smiling can be an expression of joy, blushing can be an expression of shame, and picking up an umbrella can be an expression (at least) of the desire to keep oneself dry. Examples of mental expressions like these can be characterized either *relationally* or *non-relationally*.

Relational views of mental expression consider that expressions are pieces of behavior that acquire their expressive content by virtue of their relation to a different set of items: certain internal states of the subject. This relation can be understood either *causally* or *mereologically*. Relational views of a causal type consider that mental states (e.g., pain) are identical with certain neurophysiological or functional states of the subject and that expressions (e.g., crying) are causal effects of those mental states. This is the view of expression endorsed by identity theories (e.g., Place, 1995; Smart, 1995) and by functionalist theories (e.g., Armstrong, 1995; Putnam, 1995) of the mind. By contrast, relational views of a mereological type consider that mental states (e.g., fear) are a whole constituted both by expressive components (e.g., screaming, shivering, running away, etc.) and by internal non-expressive components (e.g., a sensation of fear, thoughts, visual imagery, etc.) so that expressions are related to the other internal non-expressive components of mental states as parts of the same whole or mental state. This view of expression is endorsed, for example, by behaviorism (e.g., Carnap, 1995; Hempel, 1980),[1] and more recently, by some defenders of the idea that it is possible to have direct perceptual access to other people's mental states (e.g., Bar-On, 2004; Danón & Kalpokas, 2017; Green, 2007; Krueger, 2012; Krueger & Overgaard, 2012).

Relational views of mental expression are mainstream. However, they are not mandatory. According to non-relational views, expressions are pieces of behavior whose expressive content is an intrinsic property rather than the result of a relation to a different set of items. Mental states are *patterns of expressive behavior*, where expressive behavior (e.g., tears of sadness) is considered a species of a different genus than behavior simpliciter (e.g., tears of allergy). Since mental states are conceived as patterns of expressive behavior, there can't be any items other than expressions in mental states to which expressions themselves could be related in order to acquire their expressive content. Non-relational views of expression can be traced back to Wittgenstein[2] and they seem to be well supported by his classical argument against the possibility of private languages.[3]

Non-relational views of expression have not been much explored in the contemporary debate about the mind.[4] It is probably believed that they are unable to explain those

---

[1] Behaviorism characterizes as behavior both outer expressions and neurophysiological non-expressive states internal to the subject. See, for example, Hempel (1980, p. 18).

[2] "'Grief' describes a pattern which recurs, with different variations, in the weave of our life. If a man's bodily expression of sorrow and of joy alternated, say with the ticking of a clock, here we should not have the characteristic formation of the pattern of sorrow or of the pattern of joy. 'For a second he felt violent pain.'—Why does it sound queer to say: 'For a second he felt deep grief'? [...]". (Wittgenstein, 1953, part II, i)

[3] See García Rodríguez (2020).

[4] The only contemporary defense of non-relational expressivism that I know of can be found in García Rodríguez (2018, 2019, 2020, 2021, 2023).

mental phenomena with an internal and *apparently* inexpressible aspect to them, such as self-knowledge, inner speech, visual imagery, or pretense. Take, for instance, inner speech. If inner speech consists in the occurrence of linguistic thoughts in one's mind as if one were talking silently and without pronouncing any sound, how could these thoughts be accounted for as expressions? Sure, one could report her thoughts later on with the linguistic expression "I thought such and such", but this report cannot be identified with the thoughts of inner speech because those thoughts took place at t_1 and the report is issued later on at t_2 (if it is issued at all). Then, the question remains: how could those thoughts be accounted for as expressions?

This article aims to show that, contrary to appearances, inner speech can only be appropriately explained from the framework of non-relational expressivism. The structure of the argument is as follows. Firstly, a paradigmatic example of inner speech will be described and two desiderata for accounts of inner speech will be set. Secondly, it will be argued that relational views of expression are unable to account for the two desiderata of inner speech. Thirdly, a non-relational expressivist view of mental states will be developed. Finally, it will be argued that only from the framework of this non-relational view of mental states it is possible to explain inner speech appropriately: cases of inner speech will turn out to be expressive episodes in which something is done with words (speech acts), but in which the physical sounds of those words are missing from the sets of physical components that constitute them as expressive episodes.

2. *Two Desiderata for Accounting for Inner Speech*

Cases of inner speech are cases of thinking silently in words. The variety of features that inner speech might have in different cases, and the variety of functions that it might play within a subject's psychology, are common topics of empirical investigation.[5] However, this section is going to present inner speech under a different focus. Here, a paradigmatic example of inner speech is going to be described and analyzed to find out two features of inner speech that will reveal themselves as key in the dispute to explain the phenomenon. Let's see the example:

> Tom is worried about the financial state of the company in which he works and about the possibility of losing his job. He is at the moment having lunch with his friend, Mary, talking about where they would like to travel for their next holidays. However, while Mary is telling him her plans for next summer, Tom becomes absent from the conversation and the linguistic thought "I might be unemployed next summer if the company doesn't improve its revenue" silently crosses his mind, together with a certain facial expression, bodily posture and general demeanor.

During the time that Tom is silently thinking "I might be unemployed next summer if the company doesn't improve its revenue", Tom is engaging in inner speech. Taking this example as a guide, there are two features of inner speech that are important to point out. The

[5] See Vicente & Martínez-Manrique (2011) for an overview of the debate on the nature and functions of inner speech.

first feature is that the thoughts of a subject who engages in inner speech are hidden from everyone else in the sense that nobody can know the precise thought that the subject is having if he doesn't tell later on (that's why it's called *inner* speech). Mary might notice that Tom is distracted and absent from the conversation, she might even notice that he is worried about something, but Mary can't know by herself that Tom was thinking "I might be unemployed next summer [...]" when she was telling him her holiday plans.

The second feature of inner speech is that a subject who engages in inner speech might be able to report his thoughts later on, and a qualified audience might be able to tell when that report is false, at least on favorable occasions and with a varying degree of certainty. Imagine that Mary and Tom meet again the day after to have lunch and she asks him: "What did you think yesterday when I was telling you my plans for the summer? You looked worried all of a sudden". Imagine that Tom doesn't want to tell the truth (which is that he was thinking about the possibility of losing his job) and he answers saying "I got distracted thinking about organizing a dinner party that night". In this case, Mary might be able to tell with a certain degree of certainty that Tom's report is false, that he wasn't thinking about organizing a dinner party, on the basis of him looking worried yesterday during their conversation, not showing any interest in meeting more people later that day, and so on.

What is interesting about these two features of inner speech is that they give us two safe *desiderata* for explaining inner speech. If an account fails to be compatible with any of these two features, it should be disregarded as inadequate. So, any good account of inner speech must respect the following two desiderata:

> 1) It must be compatible with the fact that thoughts are *hidden from everyone but oneself* during inner speech.
>
> 2) It must be compatible with the *possible existence of public and intersubjective criteria* to detect false first-person thought reports about previous acts of inner speech.

The necessity of the first desideratum is obvious. However, the second desideratum requires an explanation. If there are cases in which it is possible to detect that a first-person thought report is false (e.g., Mary and Tom's case), it is clear that some public and intersubjective *evidence* of the falsehood of those first-person thought reports must exist. But then, why does the second desideratum require *criteria* rather than general *evidence*?

Two types of evidence can be distinguished: *symptoms* and *criteria*.[6] Symptomatic evidence is the kind of evidence that we have when we have access to an item B through an item A that is empirically correlated with B. For example, if the needle of the gauge of my car (A) indicates that I have ten liters of fuel left in the tank (B), I have symptomatic evidence of the fact that I have ten liters of fuel left in the tank because there is an empirical correlation between what the needle says and the amount of fuel left in the tank. By contrast, criterial evidence is the kind of evidence that we have when we have direct access to an item in one of its aspects. I have criterial evidence of the fact that there are ten liters of fuel left in the tank if I extract the liquid from the tank and I measure it directly observing where it settles in a measuring container.

[6] See: Albritton, 1959; De Gaynesford, 2002; Lycan, 1971; Shoemaker, 1963; Witherspoon, 2011; Wittgenstein, 1958; Wright, 1984.

Given this distinction between symptoms and criteria, one could argue against the second desideratum that it is enough for an account of inner speech to be compatible with the possible existence of *symptomatic evidence* to detect false first-person thought reports; that criterial evidence is not needed here at all. Suppose that there is an empirical correlation between having certain kinds of expressions (e.g., a facial expression of worry, a lack of interest in meeting more people that day, etc.) and having certain kinds of thoughts (e.g., "I might be unemployed next summer") rather than others (e.g., "I should organize a dinner party tonight"). If that were the case, Mary seemingly could detect the falsehood of Tom's report "I was thinking about organizing a dinner party" using as symptomatic evidence his expressive behavior during inner speech. For there would be a lack of empirical correlation between Tom's expressive behavior (i.e., a facial expression of worry, no interest in meeting more people that day, etc.) and the thought that he reported having (i.e., "I should organize a dinner party tonight"). So, against the second desideratum, symptomatic evidence would be all that is needed to explain how it is sometimes possible to detect false first-person thought reports.

However, things cannot be as described in this objection. The second desideratum specifies criteria, rather than symptoms or general evidence, because symptomatic evidence is epistemically dependent on criterial evidence. To find out that there is an empirical correlation between the items A and B (symptomatic evidence), it is necessary to *first* experience the items A and B directly (thereby having criterial evidence of both A and B in the first place). For example, to find out that there is an empirical correlation between barometers indicating low pressure and rain so that the former event is symptomatic evidence of the latter, it is necessary to first experience directly (criterial evidence) multiple cases of barometers indicating low pressure followed by rain. If having direct access to barometers indicating low pressure or to events of rain (criterial evidence) were somehow impossible, it would also be impossible to find out that barometers indicating low pressure are empirically correlated with events of rain (symptomatic evidence). Similarly, if subjects could only rely on symptomatic evidence to detect the falsehood of first-person thought reports and criterial evidence were completely out of the equation, it would be impossible to find any empirical correlation, or lack of empirical correlation, between kinds of expressions and kinds of thoughts. For, *ex hypothesi*, no aspect of other people's thoughts could ever be directly experienced (criterial evidence). As a result, no symptomatic evidence of the falsehood of a first-person thought report could ever be found in a subject's expressive behavior.

Therefore, if cases in which it is possible to detect the falsehood of a first-person thought report on the basis of evidence are to be explained (e.g., Mary and Tom's case), the second desideratum has to specify that accounts of inner speech must be compatible with the possible existence of public and intersubjective *criteria*. Of course, sometimes it is possible to judge that a first-person thought report is false on the basis of symptomatic evidence (e.g., the testimony of a common friend). But the second desideratum must prescribe compatibility with the existence of criterial evidence because, given the epistemic dependence of symptomatic evidence on criterial evidence, symptomatic evidence cannot be the *only way* to detect false first-person thought reports.

At first glance, the two desiderata of inner speech might seem contradictory. If thoughts are hidden from everyone but oneself during inner speech, how are we supposed to have public and intersubjective criteria to detect false first-person thought reports?

Throughout this article, it will be shown that they are not contradictory and that only within the framework of non-relational expressivism is it possible to develop an account of inner speech that fulfills both.

3. Relational Views of Expression and Inner Speech

This section will argue that it is impossible to explain inner speech from the framework of a relational view of expression, whether in its causal or mereological version. Since mereological accounts of inner speech seem better equipped than causal accounts to resist the argument that is going to be developed here, this section will focus first on mereological accounts. However, the applicability of this argument to causal accounts is straightforward and it will be made clear at the end of the section.

Bar-On (Bar-On, 2004; Bar-On & Ochs, 2018) is one of the authors who tries to account for inner speech from a mereological view of mental states. She considers that the relation between a mental state and its expressions is analogous to the one between a tree and its branches (2004, p. 228). Mental states are conditions of a subject (2004, p. 424), which have expressions among their components (2004, pp. 228, 424), but which cannot be reduced to a repertoire of expressions (2004, p. 421). Since expressions are *among* the components of mental states (which cannot be reduced to a repertoire of expressions), there must be other internal non-expressive components in mental states as well (e.g., sensations, thoughts, visual imagery, etc.). From this mereological framework, Bar-On & Ochs (2018) sketch an account of those cases of inner speech whose thought is an avowal. According to Bar-On's neo-expressivist view of self-knowledge, outer avowals (e.g., saying aloud "I'd love a cup of tea") are self-ascriptions of mental states that *voice out* or *speak out* (i.e., directly express) the very same mental state that they self-ascribe (e.g., my desire to have a cup of tea). Analogously, Bar-On & Ochs consider that those cases of inner speech whose thought is an avowal (e.g., thinking silently "I'd love a cup of tea") are *"acts of innerly speaking our mind"* (2018, p. 19); acts of innerly expressing in a direct and unmediated way the very same mental state that they innerly self-ascribe (e.g., my desire to have a cup of tea).

Green (2007, 2010) is another author who talks about inner speech from a mereological understanding of mental states. He considers that basic emotions (e.g., anger, fear, happiness, etc.) are complexes constituted by a set of interrelated phenomena (2010, p. 50), among which there are both expressive and internal non-expressive components. For anger, Green points out such components as: "physiological response, behavioral disposition [...], a cognitive disposition to make judgments of a certain sort [and] a subjective character—what we think of as the 'boiling up' aspect of how anger feels" (2010, p. 50). From this mereological framework, Green describes the thoughts of inner speech as a "saying in our hearts" or "inner monologue" (2007, p. 34). This saying in our heart or inner monologue, according to Green, has an intentional object, can be targeted at someone or something, and can be expressive of something. Here is one of Green's examples:

> When Carrie thinks, 'This is me being brave and strong for Mike,' the object of her thought is a certain relation between her and Mike, but the target of her articulation of that thought is herself. She is telling herself that she is being brave and strong for Mike. In so doing she is expressing to herself a thought whose object is a relation between her and Mike. (2007, p. 34)

Taking these views of inner speech as examples, mereological accounts of inner speech could be generally described as follows. An act of inner speech is a mental condition of a subject constituted both by outer expressive components (e.g., facial expression, bodily posture, demeanor, etc.) and by other internal non-expressive components, among which there is a chain of linguistic thoughts (e.g., "I'd love a cup of tea" or "This is me being brave and strong for Mike"). These thoughts might have different features and might play different roles within different mereological accounts of inner speech. They might be "acts of innerly speaking our mind" or "inner monologues" with an object, target, and expressive character. However, regardless of how they are characterized, what is common to all mereological accounts is that they identify those thoughts with certain internal states of the subject. Particularly, if dualism is to be avoided, they have to identify those thoughts with certain neurophysiological or functional states internal to the subject. Therefore, let me refer henceforth to the general idea that acts of inner speech are conditions of a subject constituted both by outer expressions and by inner linguistic thoughts (identified with neurophysiological or functional states) as the *mereological account of inner speech*.

Is the mereological account of inner speech any good? Since mereological accounts identify thoughts with certain internal states of the subject (i.e., neurophysiological or functional states), it is clear that they explain why thoughts are hidden from everyone but oneself during inner speech (first desideratum). However, precisely because mereological accounts identify thoughts with such internal states (i.e., neurophysiological or functional states), they fail to explain how it is sometimes possible to detect other people's false first-person thought reports (second desideratum). Let's see what would happen with Tom and Mary's example if mereological accounts of inner speech were true.

Imagine that Tom answers Mary's question with the false report "While you were telling me your holidays plans? I thought 'I should organize a dinner party tonight'". Since the reference of this report is a thought that Tom supposedly had at the time (i.e., "I should organize a dinner party tonight"), if mereological accounts were true and thoughts were identical with certain internal states of the subject, the reference of this report would be one of Tom's internal states at that time (i.e., a certain neurophysiological or functional state). However, if that were the case, Mary wouldn't be able to tell that Tom's report is false (as she can actually do) because she couldn't have any criterion to judge whether or not Tom really thought "I should organize a dinner party tonight". Indeed, there are three candidates for Mary's criteria: 1) Tom's actual thought during inner speech, 2) Tom's later *true* report about what he thought, and 3) Tom's expressive behavior during inner speech. But if mereological accounts were true, none of these three candidates could be Mary's criterion to detect that Tom's report is false. Let's see why, in order.

1) Tom's actual thought during inner speech (i.e., "I might be unemployed next summer") can't be Mary's criterion to judge that his report "I was thinking about organizing a dinner party" is false because, as long as thoughts are identified with certain internal states (i.e., neurophysiological or functional states), that thought couldn't be directly accessed by Mary in any possible situation of her *day-to-day* interactions with Tom.

Of course, if Mary were to use a "brain scanner" in a laboratory, she could directly access (criteria) the neurophysiological states of Tom that are supposed to be identical with his thoughts (neurophysiological states are unlike Cartesian mental states in that regard). However, this procedure cannot explain how the falsehood of a first-person thought report is commonly detected. On the one hand, humanity has successfully identified the falsehood

of some first-person thought reports for millennia before the invention of brain scanners. On the other hand, even in today's world, people don't use brain scanners to identify the falsehood of first-person thought reports in everyday situations (as Tom and Mary's example shows).

Therefore, the thoughts of a subject during inner speech, identified with certain neurophysiological or functional states, cannot be the public and intersubjective criteria that we are looking for to explain how it is sometimes possible to detect the falsehood of other people's first-person thought reports.

2) Imagine that Tom, after lying to Mary at first, now wants to tell the truth. So, he tells Mary "Well, in fact, I was thinking that I might be unemployed next summer". This true report cannot be Mary's criterion to judge that Tom entertained the thought "I might be unemployed next summer", rather than the thought "I should organize a dinner party" (as Tom originally reported), because it is a general rule that a report cannot be its own criterion of truth. Indeed, a report might be *symptomatic evidence* of the occurrence of a fact different from the report itself (e.g., the report "It's raining" might be symptomatic evidence of the fact that it is raining), but a report cannot be the *criterion* of its own truth. Otherwise, every report would be true by default because it would be true by virtue of itself.

Therefore, Tom's true report ("I was thinking that I might be unemployed next summer") might be symptomatic evidence of the fact that he thought "I might be unemployed next summer", and hence, symptomatic evidence of the fact that his former report ("I was thinking about organizing a dinner party") is false. But none of these first-person thought reports can be a criterion of what Tom thought during inner speech. Otherwise, every first-person thought report from Tom's side would have had to be taken by Mary as true by default until Tom denies it later on in a second report, and hence, it would have been impossible for Mary to detect the falsehood of Tom's former report from the very beginning (as she can actually do).

3) Finally, it could be argued that Tom's expressive behavior during inner speech is Mary's criterion to detect the falsehood of his report. In fact, mereologists seem well-suited to argue this because they think that it is possible to have *direct perceptual access* to other people's mental states by perceiving their expressive behavior (Bar-On, 2004; Danón & Kalpokas, 2017; Green, 2007; Krueger, 2012; Krueger & Overgaard, 2012). Similar to how subjects can directly perceive a tree when they perceive one of its branches (outer component), mereologists consider that subjects can directly perceive other people's mental states when they perceive their expressive behavior (outer component). The expressive behavior of a subject is thus considered to be *transparent-to-the-subject's-condition* or mental state (Bar-On, 2004). As a result, mereologists could argue that Tom's expressive behavior is Mary's criterion of falsehood because, even if Mary can't directly perceive Tom's thoughts by perceiving his expressive behavior (for thoughts are the *internal* components of mental states), Mary could directly perceive Tom's mental states, of which thoughts are internal components, by perceiving his expressive behavior.

However, this is not enough to make Tom's expressive behavior Mary's criterion of falsehood. At least not from a mereological understanding of mental states. Mereologists consider that the external components of a mental state (i.e., certain expressive behaviors) can take place without the corresponding internal components (i.e., thoughts, sensations, visual imagery, etc.). Bar-On (2004), for instance, argues that a subject can express a mental state without actually having that mental state because of a lack of its characteristic inter-

nal components. Cases of acting or pretense show, she thinks, that a subject can express pain (e.g., moaning, shouting "It hurts!", etc.) without actually being in pain due to her lack of any sensation of pain. Since the expressive behavior of a subject is thus conceived as being *detachable* from the subject's actual mental states (i.e., a subject can express M without actually having M), it follows that any possible expressive behavior of a subject should be compatible with the subject having any possible mental state. Then, if mereological accounts were true, Tom's expressive behavior at the time of inner speech (e.g., a facial expression of worry, a distracted attitude, a tendency to spend the rest of the day alone, etc.) would be compatible with Tom having any thought of any kind (e.g., "I might be unemployed next summer" or "I should organize a dinner party"). But if Tom's expressive behavior were compatible with any thought of any kind, Tom's expressive behavior during inner speech couldn't be Mary's criterion to detect the falsehood of his first-person thought report.

Some mereological accounts might argue that, even if it is true that a subject can express a mental state without actually having that mental state when pretending or on a stage, there is an *empirical correlation* between expressing M and having M. Similar to how *most* tree branches are attached to a trunk as parts of a tree (rather than severed), *most* subjects who express a mental state are not pretending or on a stage, and so, they have the expressed mental state. However, this view implies that a subject's expressive behavior *could only be* symptomatic evidence of the falsehood of his first-person thought reports, and it was already argued that the second desideratum of inner speech requires compatibility with criterial evidence.

Therefore, mereological accounts are incompatible with the possible existence of public and intersubjective criteria to detect the falsehood of first-person thought reports, failing to explain the second desideratum of inner speech. There are three possible candidates for those criteria: 1) the actual thoughts of a subject during inner speech, 2) the subject's later *true* reports of those thoughts, and 3) the subject's expressive behavior during inner speech. But if mereological accounts were true, none of these three candidates could be the public and intersubjective criteria that the second desideratum requires.

This concludes our examination of mereological accounts. But what about the accounts of inner speech that assume a causal view of expression?[7] Causal accounts of inner speech can be generally characterized as follows. Acts of inner speech are chains of linguistic thoughts (e.g., "I might be unemployed next summer") identical with certain neurophysiological or functional states, and the expressive behavior of a subject during inner speech is a mere causal effect of those linguistic thoughts. Does the objection developed in this section apply also to causal accounts of inner speech? Causal accounts cannot be discussed here in detail, but there are reasons to think that it does. The difficulties of mereological accounts to explain the second desideratum arose from the fact that they couldn't bridge the gap between the subject's expressive behavior and his thoughts during inner speech to explain how it is sometimes possible to detect the falsehood of other people's first-person thought reports. Then, since causal accounts deepen the gap between the internal states of the subject and his expressive behavior even more (cause-effect vs. part-whole), it is expected that causal accounts will have even more difficulties than mereological accounts to explain the second desideratum of inner speech.

[7] Since functionalism is mainstream, most accounts of inner speech assume a causal view of expression.

It might be objected that some causal accounts of a functionalist type could explain the second desideratum of inner speech arguing that the thoughts of a subject are internal states (neurophysiological states) functionally individuated in terms of their *causal roles*, among which there are causal interactions with the expressive behavior of the subject. Insofar as different types of expressive behaviors would individuate different types of internal states, this functionalist account could allegedly explain how it is sometimes possible to use the expressive behavior of a subject during inner speech as criteria to detect a false first-person thought report.[8] However, this functionalist account would have to explain the following if it doesn't want to collapse into a non-relational expressivist position. It would have to explain why and how those internal states (neurophysiological states), functionally individuated in terms of their causal roles, *contribute to determining* the expressive content of the subject's behavior (as relational views of expression claim), rather than just being physical conditions that *enable* or *make possible* the subject's behavior (one cannot smile if one doesn't have a neural stimulus that triggers the movement of the relevant muscles) *without affecting* the expressive content of that behavior (as non-relational views of expression claim). I take the burden of proof to be on the functionalist's side because, *ex hypothesi*, the internal states are the ones that are supposed to be individuated by their causal interactions with the expressive behavior, and not the other way around (i.e., the expressive behavior individuated by its causal interactions with the internal states).

4. Non-Relational Expressivism and Mental States

The last section argued that the accounts of inner speech that assume a relational view of expression are unable to explain the phenomenon. The real challenge, though, is to develop an alternative that manages to explain the two desiderata of inner speech at once. This is what is going to be done in the remainder of this article. Firstly, in this section, a non-relational expressivist view of mental states will be developed. Then, in the following section, this non-relational view of mental states will be applied to the case of inner speech.

To develop a non-relational expressivist view of mental states, three ideas are going to be explained in order: the idea of mental states as *expressive patterns*, the idea of expressive patterns as a sequence of *expressive episodes*, and the idea of expressive episodes as constituted by a set of *physical components*.

4.1. Mental States as Patterns of Expressive Behavior

Non-relational expressivism considers that mental states are nothing over and above *patterns of expressive behavior* extended over time ($t_{1\text{-}n}$). Thus, a *type* of mental state (e.g., belief, intention, desire, emotion, pain, etc.) is a *type* of pattern of expressive behavior that defines a characteristic way in which a subject interacts with other people and with the surrounding world. Each type of expressive pattern (i.e., each type of mental state) can be *instantiated* in different ways by different subjects, or by the same subject on different occasions, but all the different instances of the same type of expressive pattern are characterized

[8] I am grateful to an anonymous reviewer for making me notice this possible objection.

by a *family resemblance*. So, for example, saying "I have a terrible headache" or "This headache…", being irritable, rubbing one's forehead with a certain facial expression, or seeking silence and darkness, are all typical expressive behaviors of having a headache, but none of these expressive behaviors is *essential* for instantiating the expressive pattern of headache (i.e., for having a headache). For I can have a headache without saying "I have a headache" or without rubbing my forehead.

A subject doesn't need to *continuously* express a mental state over a period of time ($t_{1\text{-}n}$) to qualify as having that mental state over that period of time ($t_{1\text{-}n}$). The idea of a pattern of expressive behavior is compatible with the existence of periods of *expressive silence*, in which the subject doesn't manifest expressive behavior of that mental state, and periods of *expressive peak*, in which the subject manifests expressive behavior of that mental state. For instance, I have the belief that the Earth goes around the Sun since I learned that at school around 25 years ago so that I instantiate its characteristic pattern of expressive behavior since then. However, not being particularly interested in astronomy, I don't usually have opportunities to express that belief. Thus, the expressive pattern of that belief is instantiated in this case in such a way that it has long periods of expressive silences (e.g., I've gone for years without paying any attention to the skies) and only a few moments of expressive peaks (e.g., once I taught basic astronomy to my little cousin). Nevertheless, I will have the belief that the Earth goes around the Sun as long as I instantiate its *pattern* of expressive behavior (i.e., as long as I have the disposition to manifest its relevant expressive behavior when the situation so requires).

This non-relational view of mental states involves a conception of *pretense* different from the one endorsed by mereologists, such as Bar-On (2004). If mental states are patterns of expressive behavior, to say "It's raining" insincerely or on a stage cannot be considered an *expression of* the belief that it is raining *without* the belief that it is raining (for there aren't internal components in mental states to explain why that subject doesn't qualify as having the belief). By contrast, non-relational expressivism considers that pretense is a *sui generis* mental state because it is a *sui generis* pattern of expressive behavior. To say "It's raining" insincerely or on a stage is expressive of the mental state of *pretending* that it is raining, and not of the *belief* that it is raining, because a subject who pretends that it is raining instantiates a pattern of expressive behavior different from a subject who actually believes that it is raining. For example, a subject who pretends that it is raining to deceive others or on a stage won't pick up an umbrella when he thinks there's nobody watching or after the play is over. Also, he might talk in a way, or with a tone, facial expression and bodily posture, slightly different from when he actually believes that it is raining. It is on the people who watch and know the subject to find out (or not) the differences in the patterns of expressive behavior that would reveal that he is pretending rather than believing.[9]

4.2. Expressive Patterns as Sequences of Expressive Episodes

A pattern of expressive behavior (mental state) is formed of a succession of different *episodes of expression*.[10] An episode of expression is a piece of behavior that expresses a mental

[9] This non-relational expressivist view of pretense is taken from García Rodríguez (2018).

[10] Although there could be expressive patterns/mental states formed of only one episode of expression. For instance, an intense but momentary pain.

state[11] by virtue of *what it does* and *how it does it*. For example, smiling at Tom while saying "Hey, so nice to see you!" is an episode of expression because it is a piece of behavior that expresses joy for meeting Tom by virtue of what it does (greeting Tom) and of how it does it (sincerely, positively and enthusiastically). It is thus an expressive episode of the pattern/mental state of joy for meeting Tom.

All expressive episodes of mental states have a *performative character*, which could be glossed (at least) in terms of *illocutionary force* and *perlocutionary force*. On the one hand, the illocutionary force of an episode of expression is what the episode does, and how it does it, when it is manifested by a subject in a given context. The illocutionary force of an episode of expression is what determines which type of mental state the episode expresses (i.e., its expressive content). Indeed, each type of expressive pattern/mental state (i.e., belief, desire, pain, joy, etc.) defines a way in which the subject interacts with others and the world. Hence, what an episode of expression does and how it does it (illocutionary force) determines which of the different ways of interacting with others and the world (i.e., belief, desire, pain, joy, etc.) the episode of expression instantiates. For example, taking the bus (what is done) with the intention of going to the university and taking an exam (how it's done) is an expressive episode of the belief that I have an exam today because it instantiates the way of interacting with others and the world characteristic of the belief that I have an exam today.

On the other hand, the perlocutionary force of an episode of expression is the effect that it has on other people when it is manifested by a subject. For example, saying "I have an exam today" has the effect of letting other people know that I have an exam today. Unlike the illocutionary force, the perlocutionary force of an episode of expression does not affect its expressive content. For one thing is the effect that the episode of expression *should* have on others by virtue of what it does and how it does it (illocutionary force), and another thing is the effect that it *actually* has on others in a particular case (perlocutionary force). My smile while walking among students in the corridor is an expression of friendliness towards them (illocutionary force), but it might be mistakenly taken as an expression of arrogance (perlocutionary force). Yet, that doesn't change the fact that friendliness is its real expressive content.

4.3. Expressive Episodes as Sets of Physical Components

An expressive episode is constituted by a set of *physical components* from the (perceivable or non-internal) body of the subject. For instance, the expressive episode of joy for meeting Tom described above is constituted by a set of physical elements of the subject's body, such as a smile-like distribution of the muscles of her face, some sounds coming out from her mouth ("Hey, so nice to see you!") and a certain distribution of her extremities (it would be odd to express joy for seeing a friend while moving one's arm as if one were to punch him). However, these physical elements only constitute an

[11] In fact, a piece of behavior can express multiple mental states at once (i.e., it can be an expressive episode of multiple mental states at once). For instance, me picking up an umbrella might express, at the same time, my *intention* to pick up an umbrella, my *belief* that it is raining and my *desire* not to get wet. For reasons of simplicity, this possibility will be ignored henceforth.

expressive episode of joy when they take place as a whole in a context rather than as elements in isolation. Similar to how only when the pieces of a car are put together on a road a new functionality arises (i.e., you can transport yourself), only when those physical components take place as a unity in a context an expressive episode arises (for only then something is done in a certain way —illocutionary force). An isolated smile-like distribution of the muscles of my face, physically identical with the one that I have when I express joy, could well be the result of a physical condition (e.g., a stroke) without any relation to a mental state whatsoever. It is only when that smile-like distribution of the muscles of my face takes place in a unitary conjunction with other physical elements of my body (e.g., certain bodily posture, certain movements, etc.) and in a certain context (e.g., I am walking and I run into a friend) that an expressive episode of joy arises (i.e., that I actually greet a friend by smiling at him).

Thus, the set of physical components that are constitutive of an expressive episode of a mental state are the *vehicles of expression* of that mental state, the *bearers* of expressive content, but only when they take place as a unity in a context. For only then, something is done in a certain way (illocutionary force) and an expressive episode arises.

5. *The Non-Relational Expressivist Account of Inner Speech*

In this section, the non-relational expressivist view of mental states is going to be applied to explain inner speech. The claim that is going to be defended here is that acts of inner speech are expressive episodes of mental states with the two following features: 1) they *do something with words* and 2) the *physical sounds* of their words are missing from the sets of physical components that constitute them as expressive episodes. In what follows, this non-relational expressivist account of inner speech is going to be developed further and it will be argued that it appropriately explains inner speech.

Two kinds of expressive episodes can be distinguished language-wise. On the one hand, there are expressive episodes of mental states in which the subject doesn't do anything with words (e.g., picking up an umbrella). On the other hand, there are expressive episodes of mental states in which the subject does something with words (e.g., asserting "It's raining"). The latter kind of expressive episodes are called *speech acts.*[12] This distinction is made at an expressive level because it separates two kinds of expressive episodes of mental states (non-linguistic and linguistic expressive episodes). In turn, at the level of the vehicles or bearers of expression, speech acts can be divided into two groups. On the one hand, those speech acts that include the physical sounds of their words among their physical constituents are called *acts of outer speech*. On the other hand, those speech acts that don't include the physical sounds of their words among their physical constituents, leaving an empty slot among them, are called *acts of inner speech*. Since acts of inner speech are speech acts in spite of being silent, one thing to notice is that the physical sounds of words are not needed in

[12] I endorse Austin's (1962) view of speech acts, according to which speech acts don't require the speaker's intention to convey information because speech acts are used to perform a large variety of actions with different purposes (e.g., praying, scolding, cheering up, promising, complaining, sentencing, etc.).

order to qualify *as doing something with words* (e.g., it is possible to make the assertion "It's raining", or the complaint "What a day!", both aloud or silently in thought).

The process by which people learn inner speech shows that acts of inner speech are *genealogically* and *conceptually* dependent on acts of outer speech. In line with Vygotsky's theory (2012), the learning process of inner speech could be described in three stages (which in reality are chronologically overlapped).[13] Firstly, subjects learn how to do things with words by pronouncing sounds aloud while interacting with others. For instance, a toddler learns how to express hunger by saying "I'm hungry" to her caregivers rather than crying. In this stage, some of the natural vehicles of expression of our mental states are replaced by linguistic, culturally learned, vehicles of expression (e.g., the sounds "I'm hungry" replace the tears and facial features of a cry in common expressive episodes of hunger). In the second stage, subjects learn how to do things with words without communicating or interacting with other people. A toddler might express hunger by saying or whispering "I'm hungry!" while playing alone in a room. Finally, in the third stage, subjects learn how to do things with words silently and without making any sound so that nobody can hear them even if they are not alone in the room. A toddler might silently think "I'm hungry!" while trying to sneak some food from the kitchen without being noticed. This is the stage of inner speech. Since acts of inner speech are genealogically and conceptually dependent on acts of outer speech in the way described, all acts of inner speech (e.g., thinking silently "I'm hungry!") have their counterpart in outer speech (e.g., saying aloud "I'm hungry!"), and vice versa.[14]

We are now in a position to explain the non-relational expressivist account of inner speech in a deeper way. An act of outer speech (e.g., saying aloud "I'm hungry!") and its counterpart in inner speech (e.g., thinking silently "I'm hungry!") are two different tokens, two different forms of presentation, of the same type of linguistic expressive episode (speech act) of a mental state (e.g., hunger). On the one hand, both are expressive episodes of the *same type* because they do the exact same thing with words (e.g., complaining that one is hungry) in the same relevant way (e.g., sincerely, intensely, etc.), and so, they have the same illocutionary force and expressive content (e.g., they are expressive episodes of hunger). On the other hand, they are *two different forms of presentation* of the same type of expressive episode because they differ in a property that is *accidental* expression-wise: in the physical components that constitute them as expressive episodes. In cases of outer speech, the sets of physical components that constitute them as expressive episodes include the physical sounds of words (e.g., the sounds "I'm hungry!"), *together* with other physical components, such as facial features, bodily distribution and bodily movements (e.g., a grimace of hunger, putting one's hands on one's belly, seeking food, etc.). By contrast, the physical

[13] See Alderson-Day & Fernyhough (2015) for more references about researchers describing this process.

[14] This view of inner speech is compatible with the *activity view* of inner speech (Fernández Castro, 2016; Martínez-Manrique & Vicente, 2015) and with the *commitment view* of speech acts applied to inner speech (Fernández Castro, 2019; Geurts, 2018). The activity view argues that inner speech cannot be understood as a representational vehicle to have conscious access to our thoughts. The commitment view argues that acts of inner speech are speech acts that can't be viewed as a way of conveying information to oneself. Thus, both consider that inner speech is "non-relational" in the sense that it doesn't require a relation between the second-order and the first-order realms of the subject. I agree. However, the aim of my account is different: characterizing the nature of inner speech within a non-relational view of expression.

sounds of words (e.g., the sounds "I'm hungry!") are missing from the sets of physical components that constitute acts of inner speech as expressive episodes, *leaving alone* other physical components, such as facial features, bodily distribution and bodily movements (e.g., a grimace of hunger, putting one's hands on one's belly, seeking food, etc.).[15]

Therefore, since the only difference between an act of outer speech and its counterpart in inner speech lies in a property that is accidental expression-wise (i.e., their physical constitution as expressive episodes), both are the same type of expressive episode of mental state presented in two different ways: with or without the physical sounds of words. All things being equal, the expressive content of the behavior of a subject who performs an act of inner speech is the same as it would have been if he had performed the corresponding act in outer speech. Since he would have done the same thing with words in both cases (same illocutionary force), he would have expressed the same mental state in both cases, albeit with two different sets of vehicles of expression. Thus, thinking is doing something with words (speech act) without pronouncing the sounds of those words.

Is this non-relational expressivist account able to explain the two desiderata of inner speech? Before tackling the first desideratum, it is necessary to describe the basics of how this version of non-relational expressivism understands consciousness and introspection.[16] I can be conscious of the fact that my legs are crossed in a way that you can't: I can rely on my sense of proprioception, but you need to see them with your eyes or touch them with your hands. Similarly, I can be conscious of my expressive episodes of headache in a way that you can't: I can rely on my sense of introspection to feel my pain, but you need to see my face, hear my complaints, or watch my behavior. Notice that here the sense of introspection is not supposed to grant me exclusive access to an alleged private realm of my subjectivity (e.g., my neurological system or my mental substance). By contrast, similar to how my sense of proprioception provides me with a special first-person mode of access to something that is publicly and intersubjectively accessible by other means (e.g., that my legs are crossed), my sense of introspection provides me with a special first-person mode of access to something that is publicly and intersubjectively accessible by other means: my current expressive episodes of headache (e.g., my grimace, my complaints, my irritability, that I'm rubbing my forehead, that I'm smiling to hide my pain from others, etc.).[17] My *feeling* of headache is the phenomenological result of this introspective mode of access that I have (but not you) over my current expressive episodes of headache, which are otherwise publicly and intersubjectively accessible by other means.[18]

---

[15] Notice that without any physical component, there couldn't be any expressive episode at all.

[16] A more detailed account of consciousness, introspection and self-knowledge will be the topic of an independent article.

[17] Notice that the targets of introspection are not the internal states of my body (e.g., a damaged or inflamed tissue). While these internal states might *cause* some of my mental states (e.g., my headache), the targets of introspection are the expressive episodes of my mental states.

[18] The feeling of headache shouldn't be confused with the mental state/expressive pattern of headache itself. Since having a headache is having a specific pattern of expressive behavior, I could have a mild headache if I have some expressive episodes of headache (e.g., a subtle grimace, being especially irritable, etc.) even if I don't have the feeling of headache because I am not introspectively conscious of them (maybe because I'm paying full attention to something else). However, once I become introspectively conscious of my expressive episodes of headache (maybe because I stopped what I was doing), the feeling of headache arises.

Then, why are thoughts hidden from everyone but oneself during inner speech (first desideratum)? Firstly, let's consider the case of outer speech. When I talk, I pay attention to the activity that I am performing with words in a context (e.g., asserting, doubting, commanding, comforting, teasing, etc.) rather than to myself. Despite this, since I am normally conscious of what I am saying, and since I don't need to *hear* the sounds of my words to be conscious of what I am saying, it is clear that I am normally conscious of my acts of outer speech through the sense of introspection, which is working in the background while I focus on what I am doing with words in a context.[19] However, things are different in the case of how others become conscious of what I say. Since other people can't apply their senses of introspection to my acts of outer speech, they have to *hear* the sounds of my words to find out what I am saying. The sounds of my words are the *primary* vehicle of expression through which others can find out what I am doing with words in cases of outer speech. Other vehicles of expression (e.g., my facial features, my bodily posture, my behavior...) usually play a secondary role in cases of outer speech.[20] Now let's turn to the case of inner speech. Since I am normally conscious of what I am doing with words through the sense of introspection rather than by hearing the sounds of my words (which are missing in cases of inner speech), my ability to be conscious of my speech acts is not reduced in cases of inner speech compared to cases of outer speech. However, in cases of inner speech, the expressive content of my speech acts is *partially* hidden from others because, insofar as the sounds of my words are missing, people are left only with secondary vehicles of expression (e.g., my facial features, bodily posture, demeanor, etc.) to judge what I may be thinking or doing silently with words.

Let's illustrate this point with Mary and Tom's example. When Tom silently thinks "I might be unemployed next summer", Mary might notice that Tom is worried about something on the basis of his secondary vehicles of expression (e.g., facial features, bodily posture, demeanor...), but there is an aspect of the expressive content of Tom's behavior that is hidden from Mary. Since the sounds "I might be unemployed next summer" are missing from Tom's expressive episodes, Mary can't find out what Tom is thinking or doing silently with words (i.e., asserting "I might be unemployed next summer") if he doesn't tell later on. By contrast, if Tom had done what he did with words aloud, adding the sounds of words (primary vehicle of expression), Mary could have perceived by herself that he was asserting "I might be unemployed next summer".

Therefore, even if acts of outer speech and their counterparts in inner speech are expressive episodes with the same illocutionary force and expressive content, their different vehicles of expression or physical constitution make a difference in the effect that they can have on others (*perlocutionary force*). Not all vehicles of expression are equally capable of showing to others the expressive content of all the different types of episodes of mental states that a subject can have. Since the sounds of words are the primary vehicle of expression of speech acts, and since the sounds of words are suppressed from the subjects' expres-

[19] This understanding of introspection would require the existence of a variety of neurological mechanisms designed to track the subject's different types of expressive episodes in their contexts (e.g., actions, bodily postures, grimaces, cries, speech acts, day-to-day activities, etc.). So understood, the sense of introspection might involve a cluster of diverse neurological mechanisms and it might share some of them with proprioception and perception.

[20] Although this secondary role can sometimes be very important; e.g., to detect that someone is lying to us.

sive episodes in cases of inner speech (leaving only secondary vehicles of expression), it is not possible to know what others are thinking or doing silently with words during inner speech if they don't tell later on.

Now that we have seen how the non-relational expressivist account explains the first desideratum of inner speech, let us examine how it explains the second desideratum: compatibility with the possible existence of public and intersubjective criteria to detect false first-person thought reports. Sometimes it is possible to use the expressive behavior of a subject as *evidence* to detect a false first-person thought report (e.g., I can use your smile as evidence that you weren't thinking about the recent loss of a loved one). But, how can it be explained that this expressive behavior is *criterial evidence* rather than symptomatic evidence? As was already argued, if relational views were true, the expressive behavior of a subject couldn't be criteria to detect the falsehood of his first-person thought reports. However, since non-relational expressivism considers that mental states are patterns of expressive behavior, things are different here. Within a non-relational expressivist framework, if I identify an aspect of the subject's behavior (e.g., a smile) as being expressive of a certain mental state (e.g., joy) and *I am not mistaken*, I have criterial evidence of that mental state because *necessarily* I've identified an aspect (e.g., joy) of the mental state/expressive pattern itself (e.g., joy at remembering that joke); i.e., *necessarily* I have identified an aspect (e.g., joy) of the total expressive content of the subject's behavior (e.g., joy at remembering that joke).

Let's see how the non-relational expressivist account explains Tom and Mary's example. Mary could detect that Tom's report "I was thinking about organizing a dinner party that night" is false, that that can't be what Tom was thinking or doing silently with words, because the assertion "I should organize a dinner party tonight" (which is an expressive episode of the mental state of wanting to organize a dinner party) *doesn't fit* Tom's expressive behavior at the time. On the one hand, it doesn't fit Tom's *expressive episodes* during inner speech because Tom looked worried and troubled about something when he got distracted from the conversation. On the other hand, it doesn't fit Tom's *patterns of expressive behavior* around that time because he didn't show any interest in meeting people other than Mary that day. Thus, Mary could detect the falsehood of Tom's first-person thought report using his expressive behavior during and around inner speech as criteria to judge that he wasn't thinking about organizing a dinner party. Of course, it is also possible that Tom manages to deceive Mary and makes her believe that he was thinking about organizing a dinner party. In this case, Mary would have failed to identify some aspects of Tom's behavior as what they really are: expressions of worry about losing his job. Induced by Tom's false report, Mary would have *mistakenly* taken Tom's behavior as being expressive of wanting to organize a dinner party rather than as being expressive of worry about losing his job. In other words, she would have *mistakenly* taken Tom's behavior as criteria of the truth of his report rather than as criteria of its falsehood.

Relational accounts of inner speech cannot explain the second desideratum in this way because, insofar as they consider that a mental state and its characteristic expressions can take place separately (i.e., a subject can express M without having M), they imply that the expressive content of the behavior of a subject during inner speech *doesn't have to be determined* by what he is thinking or doing silently with words. By contrast, since non-relational expressivism considers that the expressive content of the behavior of a subject is always determined by what he does and by how he does it (illocutionary force), it follows that the expressive content of the behavior of a subject during inner speech *is necessarily determined*

by what he is thinking or doing silently with words (i.e., it is impossible both to express M without having M, and to have an episode of M without expressing M). Therefore, only the non-relational expressivist account of inner speech explains why the expressive behavior of a subject during inner speech can be public and intersubjective criteria to detect the falsehood of his first-person thought reports.

Finally, first-person thought reports about previous acts of inner speech are *proposals to interpret* the expressive content of one's own behavior during inner speech in a particular way: as expressive episodes in which such and such was thought or done silently with words. Tom's report "I thought that I might be unemployed next summer" is the right interpretation (self-knowledge) of the expressive content of his own behavior during inner speech because, as a matter of fact, that was precisely what he was thinking or doing silently with words at the time of inner speech. Thus, it is explained why first-person thought reports can be *symptomatic evidence*, even if they can't be criterial evidence, of what a subject thought during previous acts of inner speech. An analogy might be useful to understand the different epistemic roles of a subject's expressive behavior (criteria) and his first-person thought reports (symptoms). Similar to how it is possible to point out with a finger (symptom) the piece of the puzzle that one considers is missing from an area, a first-person thought report points out (symptom) the thought that one considers one had during a previous act of inner speech. And similar to how whether the selected piece of the puzzle is the right one or not depends only on whether it fits within the surrounding pieces and within the general picture of the puzzle (criteria), whether the selected thought is the right one or not depends only on whether it fits within the subject's expressive episodes and within the subject's expressive patterns during and around inner speech (criteria).

Therefore, only the non-relational expressivist account manages to explain the two desiderata of inner speech. On the one hand, it explains why thoughts are hidden from everyone but oneself during inner speech because it claims that acts of inner speech are speech acts in which the physical sounds of words are missing. On the other hand, it explains the possible existence of public and intersubjective criteria to detect the falsehood of first-person thought reports about previous acts of inner speech because it claims that first-person thought reports are interpretative proposals of what one was thinking or doing silently with words during inner speech whose truth-value needs to be assessed under the criteria of one's own expressive behavior at the time.

6. Conclusion

It has been argued that only non-relational expressivism can appropriately explain inner speech. Acts of inner speech are expressive episodes of mental states in which something is done with words (speech acts), but in which the sounds of words are missing from the sets of physical components that constitute them as expressive episodes. The thoughts of a subject who engages in inner speech are hidden from others in his expressive behavior because the sounds of words are missing so that nobody can hear the speech acts. That hidden expressive content, though, can be communicated to others later on with a true first-person thought report (symptomatic evidence). First-person thought reports about previous acts of inner speech are interpretative proposals of what one thought or did silently with words during inner speech. When those first-person thought reports are false, their falsehood

could be detected by others taking the relevant expressive behavior as the public and intersubjective criteria of what one thought during inner speech.

By contrast, if relational accounts of inner speech were true, it would be impossible to explain how people can detect the falsehood of first-person thought reports about previous acts of inner speech. If thoughts were internal states of the subject (i.e., neurophysiological or functional states) causally or mereologically related to the subject's expressive behavior, the expressive content of the subject's behavior during inner speech wouldn't be determined by what the subject is thinking or doing silently with words. Consequently, there couldn't be any public and intersubjective criterion to detect the falsehood of first-person thought reports about previous acts of inner speech.

This article yields some unexpected results. On the one hand, contrary to appearances, a non-relational view of expression is capable of explaining inner speech (a *prima facie* unfriendly phenomenon). On the other hand, contrary to appearances, relational views of expression fail to explain inner speech due to the conceptual implications of one of their core claims: that thoughts are internal states of the subject somehow related to the subject's expressive behavior. Since most of the historical and contemporary views of the mind assume a relational view of expression from the outset (e.g., identity theories, functionalist theories, behaviorism, dualism, etc.), this conclusion has significant consequences. If the argument presented in this article is sound, mental states are patterns of expressive behavior empirically correlated with certain neurophysiological states internal to the subject (whether animal or human), but mental states cannot be identified with those neurophysiological states or with their functions.

Finally, even though non-relational expressivism identifies mental states with patterns of expressive behavior, unlike behaviorism, it does not deny that we have a rich internal life full of experiences resulting from being introspectively conscious of our own mental states. This internal life, however, is not the result of having exclusive first-person access to the alleged private realm of our subjectivity where mental states are supposed to be (as if we had mental states first and then we learned how to express them). Instead, mental states/expressive patterns are publicly and intersubjectively accessible by nature, and our internal life is the result of learning how to suppress the relevant vehicles of expression to hide the expressive content of our behavior from others (but not from ourselves).

REFERENCES

Albritton, R. (1959). On Wittgenstein's use of the Term "Criterion". *The Journal of Philosophy*, *56*(22), 845-857.

Alderson-Day, B., & Fernyhough, C. (2015). Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychological Bulletin*, *141*(5), 931-965.

Armstrong, D. (1995). The Causal Theory of the Mind. In Lyons, W. (Ed.), *Modern Philosophy of Mind*. London: Everyman, 175-190.

Austin, J. (1962). *How to do Things with Words*. Cambridge (Mass.): Harvard University Press.

Bar-On, D. (2004). *Speaking my Mind: Expression and Self-knowledge*. Oxford: Oxford University Press.

Bar-On, D., & Ochs, J. (2018). The Role of Inner Speech in Self-Knowledge: Against Neo-Rylean Views. *Teorema: Revista Internacional de Filosofía*, *37*(1), 5-22.

Carnap, R. (1995). Psychology in the Language of Physics (1931). Lyons, W. (ed.), *Modern Philosophy of Mind*. London: Everyman, 175-190.

Danón, L. & Kalpokas, D. (2017). Perceiving Mental States: Co-presence and Constitution. *Filosofia Unisinos*, *18*(2), 87-97.
De Gaynesford, M. (2002). Blue Book Ways of Telling: Criteria, Openness and Other Minds. *Philosophical Investigations*, *25*(4), 319-330.
Fernández Castro, V. (2016). Inner Speech in Action. *Pragmatics & Cognition*, *23*(2), 238–258.
Fernández Castro, V. (2019). Inner Speech and Metacognition: A Commitment-based Approach. *Logos & Episteme*, *10*(3), 245-261.
García Rodríguez, Á. (2018). Direct Perceptual Access to Other Minds. *International Journal of Philosophical Studies*, *26*(1), 24-39.
García Rodríguez, Á. (2019). Expression and the Transparency of Belief. *European Journal of Philosophy*, *27*(1), 136-147.
García Rodríguez, Á. (2020). A Wittgensteinian View of Mind and Self-knowledge. *Philosophia*, *48*(3), 993-1013.
García Rodríguez, Á. (2021). How Emotions are Perceived. *Synthese*, *199*(3-4), 9433-9461.
García Rodríguez, Á. (2023). *El pensamiento de los animales: Un modelo expresivo*. Madrid: Cátedra.
Geurts, B. (2018). Making Sense of Self Talk. *Review of Philosophy and Psychology*, *9*(2), 271-285.
Green, M. S. (2007). *Self-Expression*. Oxford: Oxford University Press
Green, M. S. (2010). Perceiving Emotions. *Proceedings of the Aristotelian Society*, *84*(1), 45-61.
Hempel, C. G. (1980). The Logical Analysis of Psychology. Block, N. (ed.), *Readings in Philosophy of Psychology* (vol. 1). Cambridge, MA: Harvard University Press, 14-23.
Krueger, J. (2012). Seeing Mind in Action. *Phenomenology and the Cognitive Sciences*, 11, 149-173.
Krueger, J., & Overgaard, S. (2012). Seeing Subjectivity: Defending a Perceptual Account of Other Minds. *Consciousness and Subjectivity*, *47*, 297-319.
Lycan, W. G. (1971). Noninductive Evidence: Recent Work on Wittgenstein's "Criteria". *American Philosophical Quarterly*, *8*(2), 109-125.
Martínez-Manrique, F., & Vicente, A. (2015). The Activity View of Inner Speech. *Frontiers in Psychology*, *6*, 232.
Place, U. T. (1995). Is Consciousness a Brain Process?. Lyons, W. (ed.), *Modern Philosophy of Mind*. London: Everyman, 106-116.
Putnam, H. (1995). Philosophy and Our Mental Life. Lyons, W. (ed.), *Modern Philosophy of Mind*. London: Everyman, 133-147.
Shoemaker, S. (1963). *Self-Knowledge and Self-Identity*. Ithaca, NY: Cornell University Press.
Smart, J. J. C. (1995). Sensations and Brain Processes. Lyons, W. (ed.), *Modern Philosophy of Mind*. London: Everyman, 117-132.
Vicente, A., & Martínez-Manrique, F. (2011). Inner Speech: Nature and Functions. *Philosophy Compass*, *6*(3), 209-219.
Vygotsky, V. (2012). *Thought and Language*. Cambridge, MA: MIT Press.
Witherspoon, E. (2011). Wittgenstein on Criteria and The Problem of Other Minds. In Kuusela, O. & McGinn, M. (Eds.), *The Oxford Handbook of Wittgenstein*. Oxford: Oxford University Press.
Wittgenstein, L. (1953). *Philosophical Investigation*. (G. E. M. Anscombe, Trans.). New York: Macmillan Publishing Co.
Wittgenstein, L. (1958). *The Blue and Brown Books*. Oxford: Blackwell.
Wright, C. (1984). Second Thoughts about Criteria. *Synthese*, *58*(3), 383-405.

Jesús López Campillo holds a PhD in Philosophy from the University of Murcia. He is interested in philosophy of mind, epistemology and philosophy of perception.

Address: Philosophy Department, Faculty of Philosophy, University of Murcia, Avda. Teniente Flomesta, N.º 5, Ed. Convalecencia, 30003, Murcia, Spain. E-mail: jesuslcampillo@gmail.com – ORCID: https://orcid.org/0000-0003-0216-5637