UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

The Transparency of Belief and The First-Person Perspective

La Transparencia de la Creencia y la Perspectiva de Primera Persona

D. Jesús López Campillo

2020

PhD Supervisor:

Ángel María García Rodríguez

A mis padres, a mis abuelos y
abuelas, y a mi tía Fina

Table of Contents

Resumen

El principal argumento de este trabajo es que la mejor explicación del autoengaño y de la paradoja de Moore se sigue de una interpretación expresivista-conductual de la Transparencia de la creencia.

La Transparencia de la creencia (ej., Evans, 1982) consiste en el hecho de que la pregunta "¿Crees que p?" se contesta algunas veces usando el mismo procedimiento que la pregunta "¿Es p el caso?" (de modo que la primera es "transparente" a la segunda). En lugar de inspeccionándose a uno mismo para rastrar si se tiene la creencia de que p (como sugiere la pregunta "¿Crees que p?"), algunas veces es posible contestar mirando al mundo y deliberando sobre las razones a favor y en contra del hecho p. Las explicaciones epistémicas de la Transparencia (ej., Fernández, 2013; Moran, 2001) afirman que la pregunta "¿Crees que p?" versa sobre la creencia p del sujeto y que el hecho de que algunas veces se conteste de la misma forma que la pregunta "¿Es p el caso?" se debe a que se ha aplicado el proceso de formación de creencias característico del autoconocimiento de primera persona (i.e., creencia verdadera y justificada). Por el contrario, la explicación expresivista-conductual de la Transparencia (García, 2019a) que se va a defender en este trabajo afirma que la pregunta "¿Crees que p?" puede preguntarse tanto en un sentido deliberativo como en un sentido autoadscriptivo. Por un lado, cuando se pregunta en un sentido deliberativo, la pregunta es sobre el hecho p y es contestada de la misma forma que "¿Es p el caso?" porque es contestada con un juicio sobre p; i.e., con un episodio expresivo de creencia. Es en este caso cuando surge el fenómeno de la Transparencia (i.e., la pregunta "¿Crees que p?" es transparente a "¿Es p el caso?") porque la pregunta se responde aquí desde el punto de vista deliberativo de la primera persona. Por otro lado, cuando se pregunta en un sentido autoadscriptivo, la pregunta es sobre la creencia p del sujeto y es contestada de distinta forma que "¿Es p el caso?" porque es contestada por autoinspección en base a evidencia (ej., emociones, acciones, juicios, pensamientos, etc.) sobre la propria creencia p y no en base a una deliberación sobre el hecho p. En este caso, el fenómeno de la Transparencia no aparece (i.e., la pregunta "¿Crees que p?" no es transparente a la pregunta "¿Es p el caso?") porque la pregunta es contestada desde el punto de vista de la tercera persona (i.e., mediante autoinspección de los propios estados mentales en base a evidencia).

Contra las concepciones epistémicas de la Transparencia, la concepción expresivista-conductual afirma que el autoconocimiento puede ser entendido de forma epistémica (*saber qué*) o de forma no epistémica (*saber cómo*). Por un lado, el autoconocimiento en sentido epistémico (*saber qué*) consiste en tener una creencia verdadera y justificada sobre los propios estados mentales y la concepción expresivista conductual considera que es un fenómeno exclusivo de la tercera persona, de modo que los sujetos sólo pueden adquirir autoconocimiento en sentido epistémico (i.e., creencia verdadera y justificada) cuando se autoinspeccionan a sí mismos usando evidencia sobre sus propios estados mentales. Por otro lado, el autoconocimiento en sentido expresivo (*saber cómo*) es entendido por la concepción expresivista conductual como expresión autoconsciente y es considerado el único sentido en el que puede hablarse de autoconocimiento de primera persona. Un sujeto se expresa autoconscientemente (y, por tanto, tiene autoconocimiento en sentido expresivo) cuando expresa un estado mental sabiendo la habilidad que está ejercitando, es decir, sabiendo lo que se trae entre manos. Por ejemplo, mi acción de coger el paraguas es autoconsciente si realizo la acción de coger el paraguas atentamente y sabiendo lo que estoy haciendo (ej., coger el paraguas para llevarlo conmigo fuera de casa). Mientras que un sujeto se expresa no-autoconscientemente (y, por tanto, carece de autoconocimiento en sentido expresivo) cuando expresa un estado mental sin saber la habilidad que está ejercitando, es decir, sin saber lo que se trae entre menos. Por ejemplo, mi acción de coger el paraguas es no-autoconsciente si lo cojo distraído y sin saber lo que me traigo entre manos, sólo para descubrir más tarde, al abrir la mochila y encontrarlo dentro, que de hecho cogí el paraguas sin darme cuenta al salir de casa.

Las concepciones epistémicas de la Transparencia asumen una concepción *relacional* de la expresión según la cual un estado mental (ej., creencia) y su conjunto de expresiones características (ej., decir "Creo que llueve", coger el paraguas, etc.) son dos conjuntos de ítems distintos que establecen algún tipo de relación entre sí. Dado que según esta concepción relacional de la expresión un estado mental es un ítem distinto de su conjunto de expresiones características, la concepción relacional de la expresión es coherente con la idea de autoconocimiento de primera persona en sentido epistémico (i.e., creencia verdadera y justificada) porque es coherente con la idea de que los estados mentales son ítems internos (en el sentido de que van más allá de lo que puede ser percibido, es decir, más allá de la expresión) que son accesibles para el sujeto mediante un procedimiento especial que solo puede usarse en el propio caso (i.e., el procedimiento de formación de creencias característico del autoconocimiento de primera persona en sentido epistémico). Así, las concepciones epistémicas de la Transparencia consideran que el autoconocimiento de primera persona en sentido epistémico (i.e., creencia verdadera justificada) es un fenómeno real que se explica por la existencia de un procedimiento especial por el cual los sujetos pueden conocer en exclusiva sus propios estados mentales. Sin embargo, en la medida en la que la concepción expresivista conductual de la Transparencia considera que el autoconocimiento de primera persona en sentido epistémico (i.e., creencia verdadera y justificada) no existe, la concepción expresivista conductual adopta una concepción *no relacional* de la expresión (García, 2018) según la cual los estados mentales (ej., creencia) son idénticos a patrones expresivos extendidos a lo largo del tiempo (ej., decir "Creo que llueve", coger el paraguas, etc.). Así, dado que los estados mentales no son ítems privados diferentes de su conjunto de expresiones características (sino que son patrones expresivos que se extienden a lo largo del tiempo), la idea de autoconocimiento de primera persona en sentido epistémico no tiene cabida dentro de la concepción expresivista conductual de la Transparencia (pues la idea de autoconocimiento de primera persona conlleva la idea de un proceso de formación de creencias especial que da lugar a un acceso exclusivo a ese ítem o estado mental privado).

A partir de este modelo expresivista-conductual de la Transparencia de la creencia, se ofrece una explicación del autoengaño y de la paradoja de Moore. Los sujetos están autoengañados respecto a un hecho del mundo cuando manifiestan un conflicto irracional entre lo que afirman sinceramente (ej., "Creo que estoy sano") y la forma en la que actúan (ej., saltarse sus citas con el médico, evitar conversaciones sobre temas médicos, etc.). Las explicaciones del autoengaño disponibles en la literatura se dividen en dos tipos: las procedimentales y las no-procedimentales. Las procedimentales consideran que el autoengaño es el resultado del proceso por el cual se genera una creencia, un deseo o una intención, y las no-procedimentales consideran que el autoengaño consiste en un estado mental *sui generis*. Entre las explicaciones procedimentales, se encuentran las intencionalistas (ej., Pears, 1984), las motivacionalistas (ej., Mele, 2001) y las epistémicas (ej., Fernández, 2013). Las explicaciones intencionalistas consideran que el autoengaño es el resultado de la intención del sujeto de engañarse a sí mismo formando una creencia que sabe que es falsa. Las explicaciones motivacionalistas consideran que el autoengaño es el resultado de una creencia que está formada por un sesgo motivacional. Y, por último, las explicaciones epistémicas consideran que el autoengaño es el resultado de una creencia de segundo orden falsa causada por un error epistémico en el proceso de formación de creencias característico del autoconocimiento

epistémico de primera persona. Entre las explicaciones no-procedimentales (las cuales defienden que el autoengaño es un estado mental *sui generis*), hay quien defiende que el autoengaño es un estado mental de fingimiento (Gendler, 2010), un estado mental de afirmación sincera sin creencia (Audi, 1989) o un estado mental consistente en evitar un pensamiento en cuyo contenido se cree (Bach, 1981).

En este trabajo se argumenta que el autoengaño se explica desde la concepción expresivista-conductual afirmando que es un estado mental *sui generis* que conlleva tanto falta de autoconocimiento expresivo de primera persona como dificultades para adquirir autoconocimiento epistémico de tercera persona. Por un lado, el autoengaño es un estado mental que conlleva falta de autoconocimiento expresivo porque los estados mentales de autoengaño son imposibles de expresar de forma autoconsciente (i.e., sabiendo lo que uno se trae entre manos). Por ejemplo, un sujeto que esté autoengañado respecto al hecho de que está sano dirá "Creo que no tengo ningún problema de salud" al ser preguntado por su salud. Esta afirmación es una expresión de su estado mental de autoengaño, pero es una expresión no-autoconsciente, pues el sujeto cree que está ejercitando su habilidad para expresar su creencia de que está sano cuando en realidad está ejercitando su habilidad para expresar su estado mental autoengaño. Así, el sujeto carece de autoconocimiento en sentido expresivo. Por otro lado, el autoengaño es un estado mental del que resulta difícil adquirir autoconocimiento epistémico de tercera persona (i.e., creencia verdadera y justificada) porque los estados mentales de autoengaño tienen un patrón expresivo que es similar en apariencia a otros estados mentales conscientes (ej., creencia, deseo o intención), de modo que pueden ser fácilmente confundidos con estos estados mentales cuando el sujeto autoengañado se autoinspecciona a sí mismo con la intención de conocer cuáles son sus estados mentales.

La paradoja de Moore surge porque las oraciones "p, pero no creo que p" y "p, pero creo que no-p" son irracionales de afirmar a pesar de que parece que pueden ser verdaderas (pues p puede ser el caso y, al mismo tiempo, yo no creer que p —ignorancia— o yo creer que no-p —error—). Existen en la literatura cuatro tipos de explicación de la paradoja de Moore. Las explicaciones pragmáticas (ej., Rosenthal, 2005) consideran que la paradoja de Moore surge porque las oraciones de Moore carecen de condiciones de aserción. Las explicaciones psicológicas (ej., Coliva, 2016) consideran que la paradoja de Moore surge porque afirmar las oraciones de Moore conlleva algún tipo de inconsistencia psicológica. Las explicaciones epistémicas (ej., Fernández, 2013) consideran que la paradoja de Moore surge porque afirmar las oraciones de Moore solo puede ser el resultado de un error epistémico en el proceso de

formación de creencias responsable del autoconocimiento de primera persona. Y, por último, las explicaciones semánticas (ej., Heal, 1994) consideran que la paradoja de Moore surge porque las oraciones de Moore, a pesar de las apariencias, son contradictorias de afirmar.

En este trabajo se argumenta que la paradoja de Moore se explica desde la concepción expresivista-conductual de la Transparencia de la siguiente forma. Por un lado, cuando las oraciones de Moore "p, pero no creo que p" y "p, pero creo que no-p" son afirmadas desde la perspectiva deliberativa de primera persona, su afirmación es irracional porque es contradictoria. Pues desde la perspectiva deliberativa de primera persona "Creo que p" es el resultado de una deliberación sobre el hecho p, de modo que "Creo que p" es un juicio o afirmación sobre el hecho p que expresa mi creencia de que p. Por otro lado, cuando las oraciones de Moore "p, pero no creo que p" y "p, pero creo que no-p" son parcialmente afirmadas desde la perspectiva autoinspectiva de tercera persona, su afirmación no es irracional porque no es contradictoria, es decir, tiene condiciones de verdad posibles. Pues desde la perspectiva autoinspectiva de tercera persona "Creo que p" es el resultado de una autoinspección, sobre cuáles son mis creencias respecto a p, llevada a cabo en base a evidencia sobre mis propios estados mentales, de modo que "Creo que p" es un juicio o afirmación sobre mi creencia p que expresa mi creencia de segundo orden de que creo que p.

En conclusión, se argumentará que la concepción expresivista-conductual de la transparencia es la correcta porque es capaz de ofrecer la mejor explicación del autoengaño y de la paradoja de Moore entre las disponibles actualmente.

Acknowledgements

This PhD dissertation is the result of more than five years of work. I would like to thank some of the people that helped me along this long, and sometimes difficult, process. Firstly, I'd like to thank my PhD supervisor, Ángel García Rodríguez, for helping me from the beginning to the end of this work. He introduced me to the expressivist ideas about the mind, self-knowledge and Transparency back when I was an undergraduate student, and he helped me all along the different stages of this project: from the design of the research program to the final draft of this text. Pretty much all the sections of this dissertation have been enhanced with his comments and discussions. Sometimes these discussions ended up in agreement and other times in disagreement, but the final claims of this PhD dissertation were always improved as a result.

Also, I would like to thank Sarah Sawyer for helping me to understand content externalism and McKinsey paradox by discussing these topics with me thoroughly. Unfortunately, content externalism and McKinsey Paradox didn't end up making their way to the fifth chapter of this PhD dissertation, but the ideas about these topics that I developed during my period as a visiting student at the University of Sussex shaped my conception of criterial evidence, which plays a fundamental role in the argument presented here. I'm also grateful to David Finkelstein for the insightful discussions about self-knowledge and expression that we had during the period that I was a visiting student at the University of Chicago. These conversations helped me to develop some of the ideas presented in the second chapter of this PhD dissertation. Also, I'd like to thank José Zalabardo for discussing Transparency with me both at the UCL and around the different workshops and conferences that we end up meeting in Spain.

I'm also grateful to the Ministry of Economy and Competitiveness for funding this PhD with an FPI scholarship (BES-2013-064774) in the context of the research project "Self-knowledge, Expression and Transparency" (FFI2012-38908-C02-02). These funds made it possible that I could do a PhD in Philosophy. They gave me four years of resources and professional stability that allowed me to engage heart and soul in the exercise of Philosophy. I feel very grateful for this (increasingly rare) opportunity. Most of my efforts were directed at researching the issues tackled in this PhD dissertation, but I was also lucky of being able to teach some seminars on Philosophy of Mind and Philosophy of Language at the University of Murcia.

I would also like to thank my PhD colleagues (and friends) for all the philosophical discussions, support and fun that we shared along these years. Doing a PhD in Philosophy is not an easy task. To the timeless difficulties that are inherent to the exercise of Philosophy, it is currently added the additional challenge of trying to make a living out of your philosophical and teaching vocation in times in which "economic orthodoxy" still means "neo-liberal orthodoxy". However, as it is sometimes said, "Cause they were, we are; cause we are, they'll be".

Finally, I'd like to thank my family and friends for their support and generosity. I'm still surprised about how understanding they were with me when I had to lock myself at home (before locking yourself at home started to be a worldwide trend) to finish this project on time.

Introduction

Sometimes subjects answer the question “Do you believe that p?” as if they were answering the question “Is p the case?”: deliberating on the basis of epistemic grounds or reasons about *whether p* instead of on the basis of evidence about whether one *believes* that p (as the question “Do you believe that p?” might suggest). For instance, sometimes to answer the question “Do you believe that there is going to be a third world war?” I have to do exactly the same that I would have to do to answer the question “Is there going to be a third world war?”, namely, I have to deliberate about the possibility of a third world war on the basis of epistemic grounds or reasons, make up my mind, and answer the question. This phenomenon is called *Transparency of belief* (Evans, 1982) and it arises only when the subject answers the question “Do you believe that p?” from the first-person perspective.

Indeed, it is considered that subjects can answer the question “Do you believe that p?” either from the *first-person deliberative perspective* or from the *third-person self-inspective perspective*. The distinction between the first-person deliberative perspective and the third-person self-inspective perspective has to do with the kind of evidence that subjects use to answer the question “Do you believe that p?”. On the one hand, a subject answers the question “Do you believe that p?” from the first-person deliberative perspective when, in line with Transparency, she answers the question on the basis of grounds or reasons about *whether p* and not on the basis of evidence about whether she currently believes that p. For instance, I answer the question “Do you believe that there is going to be a third world war?” from the first-person perspective when I answer on the basis of grounds or reasons about the possibility of a third world war. On the other hand, a subject answers the question “Do you believe that p?” from the third-person self-inspective perspective when, against Transparency, she answers the question on the basis of evidence about *whether she currently believes that p* rather than on the basis of evidence about whether p. For instance, I answer the question “Do you believe that there is going to be a third world war?” from the third-person self-inspective perspective when I answer on the basis of evidence about what I believe; e.g., that I remember that I explained in the past why a third world war is likely, that I support an increase in the military budget, that I am planning to build a bunker in my basement, and so on.

Avowals are first-person present-tense utterances that explicitly mention a mental state; e.g., “I believe that it is raining”, “I want an ice cream”, “I intend to open the door”, “I feel terrible about what happened”, “I have a headache”, etc. Avowals can be issued from the first-person deliberative perspective (henceforth, *first-person avowals*) or from the third-person self-inspective perspective (henceforth, *third-person avowals*). When avowals are issued from the first-person deliberative perspective (i.e., first-person avowals), they are *groundless* and *authoritative* in regard to the mental state explicitly mentioned in the avowal. On the one hand, first-person avowals are *groundless* in regard to the explicitly mentioned mental state because first-person avowals are issued on the basis of no evidence about the subject’s mental states. In line with Transparency, if any evidence is taken into account, it is evidence about whether p, and not about the subject’s mental states. For instance, my first-person avowal “I believe that a third world war is likely” is groundless in regard to the explicitly mentioned mental state (i.e., belief) because it is made on the basis of no evidence about my belief but on the basis of evidence about the possibility of a third world war. On the other hand, first-person avowals are *authoritative* because they enjoy a certain presumption of truth in regard to the explicitly mentioned mental state: it is not usually questioned that the subject has the mental state that she explicitly mentions in the avowal. For instance, my first-person avowal “I believe that a third world war is likely” is authoritative in regard to the explicitly mentioned mental state (i.e., belief) because, even if it can be questioned whether a third-world war is likely or not as a matter of fact, it is not usually questioned that *I believe* that a third world war is likely when I issue the first-person avowal “I believe that a third world war is likely”.

However, things are different when avowals are issued from the third-person self-inspective perspective (i.e., third-person avowals). Third-person avowals are *not* groundless in regard to the mental state explicitly mentioned in the utterance and it is usually considered that they *cannot* be authoritative either. On the one hand, third-person avowals are not groundless because they are made on the basis of evidence about the subject's mental states. For instance, my third-person avowal "I believe that a third world war is likely" is not groundless because it is made on the basis of evidence about what I believe (e.g., that I said so to others in the past, that I support an increase in the military budget, or that I am building a bunker). On the other hand, it is usually considered that third-person avowals cannot be authoritative or presumably true because it is usually considered that subjects are in the same epistemic situation to make judgements about their own mental states on the basis of evidence about their mental states (e.g., their behaviour) than any other third-person subject (e.g., a relative or a friend). For instance, if I issue the third-person avowal "I believe that a third world war is likely" only on the basis of the fact that I remember me explaining so to my cousin last year, my avowal is not authoritative or presumably true because it can be easily questioned by someone else: my cousin can deny that that was the point that I was making during the conversation (my memory can fail) or he could argue that I changed my mind six months later in another conversation that we had about the issue and that I don't remember anymore.

It is certainly true that, in the latter case, the evidence on the basis of which I issue my third-person avowal is so weak that it is not authoritative or presumably true whatsoever. However, as it will be argued in due time, it is a mistake to consider that third-person avowals can't ever be authoritative or presumably true. Third-person avowals are authoritative *sometimes* because subjects can sometimes make judgements about their own mental states on the basis of more and better evidence about their mental states than the evidence that could possibly be available to third-person subjects in their particular third-person situations. For instance, if I issue the third-person avowal "I believe that a third world war is likely" on the basis of more and better evidence about what I believe than the evidence that can be available to third-person subjects (e.g., that I remember myself silently deliberating about the issue last week until I ended up convinced, that I remember myself building a bunker in my basement while thinking "This will come in handy in the next decade", etc.), my third-person avowal will be authoritative or presumably true in the epistemic sense that it will be warranted to a higher degree (i.e., with more and better evidence) than other people's judgements about whether I believe that a third world war is likely can possibly be, and so, it will be more difficult to challenge its truth.

Most of the accounts of Transparency available in the literature (e.g., Boyle, 2009, 2011, 2015; Byrne, 2005, 2011, 2018; Evans, 1982; Fernández, 2013; Gallois, 1996; Moran, 2001, 2003) understand Transparency and first-person avowals as *epistemic phenomena*. For they understand that both behind Transparency and behind the groundless and authoritative character of first-person avowals is the special first-person procedure that subjects allegedly use to acquire *first-person epistemic self-knowledge*: true strongly[1] warranted beliefs about one's own mental states. On the one hand, in regard to Transparency, epistemic accounts consider that the question "Do you believe that p?" asks about the *subject's beliefs* both when the question is meant in a deliberative way (i.e., meant to be answered by first-person deliberation about whether p) and when the question is meant in a self-ascriptive way (i.e., meant to be answered by third-person self-inspection). Then, the question "Do you believe that p?" is considered to be semantically different from the question "Is p the case?" both when it is meant in a deliberative and in a self-ascriptive way. However, only when the question "Do you believe that p?" is meant in a deliberative way, it is supposed to be answered in the same way as the question "Is p the case?" because only in this case the subject is supposed to apply the special first-person procedure responsible for first-person epistemic self-knowledge. For regardless of the particular way in which this special first-person procedure is characterized by the different epistemic accounts of Transparency, it is always considered that it involves the subject's deliberation on the basis of evidence about *whether p* rather than about whether *she believes that p*.

On the other hand, in regard to first-person avowals, which are supposed to answer the deliberative question "Do you believe that p?", epistemic accounts of Transparency think that they are groundless and authoritative self-ascriptions of mental states because they are also the result of applying the special procedure responsible for first-person epistemic self-knowledge. For once again, regardless of the particular way in which this special first-person procedure is characterized by the different epistemic accounts of Transparency, it is supposed to deliver true beliefs about one's own mental states on the basis of evidence about *whether p* rather than on the basis of evidence about whether one believes that p (i.e., it is groundless) and it is supposed to warrant those beliefs in a stronger way than the beliefs of other people about one's own mental states can possibly be because it is supposed to be more reliable and less prone to error (i.e., it is authoritative) than other procedures of belief-formation, such as perception or inference. As a result, epistemic accounts of Transparency consider that subjects can acquire first-person epistemic self-knowledge by applying a special first-person groundless and authoritative procedure of belief-formation and that subjects can express that first-person

[1] "Strongly" because they are supposed to be better warranted than other people's beliefs about one's own mental states can possibly be.

epistemic self-knowledge with first-person avowals (i.e., groundless and authoritative self-ascriptions of mental states).

Against epistemic accounts of Transparency, a new account of Transparency based on the concept of expression has been recently proposed (García, 2019a). According to this expressivist account, the question "Do you believe that p" has different meanings when it is meant in a deliberative and when it is meant in a self-ascriptive way. On the one hand, when the question "Do you believe that p?" is meant in a deliberative way, it asks about *whether p*, and so, it is transparent to the question "Is p the case?" because both questions have the same meaning. As a result, the answer to the question "Do you believe that p?" meant in a deliberative way is a *judgement about whether p* (rather than a self-ascription of attitude) that can be expressed in the form of a first-person avowal (e.g., "I believe that p") or in the form of an assertion (e.g., "p is the case") and that is made by first-person deliberation on the basis of evidence about whether p. On the other hand, when the question "Do you believe that p?" is meant in a self-ascriptive way, it asks about the subject's beliefs, and so, it is not transparent to the question "Is p the case?" because both questions have different meanings. As a result, the answer to the question "Do you believe that p?" meant in a self-ascriptive way is a *self-ascription of attitude* that can be expressed in the form of a third-person avowal (e.g., "I believe that p") or in the form of an assertion (e.g., "It is the case that I believe that p") and that it is made by third-person self-inspection on the basis of evidence about one's own mental states. Thus, Transparency is conceived here as a *semantic* rather than as an epistemic phenomenon: the question "Do you believe that p?" is transparent to the question "Is p the case?" when its meaning is tantamount to the meaning of the question "Is p the case?".

This semantic account of Transparency is also an *expressivist account* (García, 2018, 2019b) because it is based on an expressivist view of the nature of mental states and of first-person self-knowledge. Firstly, the semantic account of Transparency endorses an expressivist or non-relational view of expression according to which *mental states* (e.g., the belief that it is raining) and their characteristic *set of expressions* (e.g., saying "[I believe that] it is raining", picking up the umbrella, spending the evening at home if one doesn't want to get wet, etc.) are one and the same item because mental states are identical to patterns of expressive behaviour. Indeed, a mental state is nothing over and above a *temporal pattern of expression*: a set of *expressive episodes* (e.g., saying "[I believe that] it is raining", picking up the umbrella, spending the evening at home, etc.) manifested by the subject in a certain way and over a certain period of time (i.e., over which the subject has the mental state in question). Thanks to the non-relational view, the expressivist account can explain why first-person avowals (which answer the deliberative question "Do you believe that p?") are *groundless* and *authoritative*, in spite of the fact that they are considered to be judgements about whether p rather than self-ascriptions of mental states, in the following *non-epistemic* way. On the one hand, first-person avowals are groundless in the *expressive sense* that they are episodes of expression of mental states, and so, like every other episode of expression (e.g., to cry out of pain), they are made on the basis of no evidence about the subject's mental states. On the other hand, first-person avowals are authoritative or presumably true in the *expressive sense* that they make explicit the mental state of which that they are an expressive episode, and so, it can only be questioned that the subject has the mental state made explicit in the utterance if there are suspicions of abnormal circumstances (e.g., insincerity, self-deception, expressive failure, etc.). For instance, the first-person avowal "I believe that it is raining" is an expressive episode of my belief that it is raining consisting in the judgement that it is raining (rather than in a self-ascription of that belief). So, it is groundless in the expressive sense that it is not made on the basis of evidence about what I believe (in fact, it is made on the basis of evidence about the rain, as Transparency requires); and it is authoritative or presumably true in the expressive sense that it can only be questioned that the subject has the explicitly mentioned mental state by arguing that abnormal circumstances are taking place (e.g., under suspicions of insincerity, self-deception, etc.) because it is an episode of expression that makes explicit the mental state that it is an episode of.

Secondly, there are at least two senses of knowledge (Ryle, 1949): *knowing that*, which is the epistemic sense of knowledge consisting in true warranted belief, and *knowing how*, which is a non-epistemic sense of knowledge consisting in having the ability to appropriately exercise an activity (e.g., swimming). The semantic account of Transparency claims that the question "Do you believe that p?" is answered with a *judgement about whether p* when it is answered from the first-person deliberative perspective (in which Transparency arises) and with a *self-ascription of attitude* when it is answered from the third-person self-inspective perspective (in which Transparency doesn't arise). As a result, from the semantic account of Transparency follows that *epistemic self-knowledge* (i.e., true warranted belief —*knowing that*—) is a third-person phenomenon rather than a first-person phenomenon, and that *expressive self-knowledge* (i.e., the exercise of the ability —*knowing how*— to express one's own mental states in a certain way) is the first-person phenomenon that naturally replaces first-person epistemic self-knowledge. On the one hand, first-person self-knowledge is an expressive phenomenon because it has to do with the exercise of the ability (*knowing how*) to deliberate about whether p, make up one's mind, and express the newly formed attitude with a judgement about whether p that answers the deliberative question "Do you believe that p?". As it will be shown, a subject has first-person expressive self-knowledge or not depending on whether she has the ability to *appropriately* express her attitude and on whether the judgement is a *self-conscious* expressive episode or not. On the other hand, since the answer to the non-transparent question "Do you believe that p?" is considered to be a self-ascription of attitude made from the third-person perspective of self-inspection, epistemic self-knowledge (i.e., true warranted belief about one's own mental states) is an only third-person phenomenon because only from the third-person self-inspective perspective there is a self-ascription of attitude, and so, only from the third-person self-inspective perspective the subject can form a true warranted belief about her own mental states (*knowing that*). This version of expressivism, which follows from the semantic account of Transparency, is going to be called *behavioural expressivism* henceforth.

The concept of expression has been recently used as well by a group of accounts called *neo-expressivist* (Bar-on, 2004, 2013; Finkelstein, 2003). Neo-expressivist accounts have in common with epistemic accounts of Transparency that they consider that first-person avowals are self-ascriptions of mental states, and so, that first-person self-knowledge is *groundless* and *authoritative* in the epistemic sense (i.e., true strongly warranted beliefs about one's own mental states based on no specific evidence about one's own mental states). According to neo-expressivist accounts, first-person avowals are groundless and authoritative because they are self-ascriptions of mental states that *express* the very same mental state that they self-ascribe (rather than expressing *only* the relevant second-order belief). For instance, the first-person avowal "I believe that it is raining" is a self-ascription of the belief that it is raining that expresses my belief that it is raining itself. Since it is an expression of the very same mental state that it self-ascribes, it is in a continuum with other expressions of mental states (e.g., a cry of pain or a smile of happiness) that are made on the basis of no specific evidence about one's own mental states (i.e., it is groundless). Also, since it is an expression of the very same mental state that it self-ascribes, the self-ascription enjoys a certain presumption of truth (i.e., it is authoritative) because to deny the truth of the self-ascription would involve the idea that some kind of expressive failure occurred so that the first-person avowal doesn't express the mental state that it self-ascribes. As a result, when subjects express their mental states with first-person avowals, they are supposed to acquire first-person epistemic self-knowledge because they are supposed to end up with true second-order beliefs strongly warranted thanks to the fact that first-person avowals express the self-ascribed mental state itself on the basis of no specific evidence about the subject's mental states. Furthermore, since the mental state of which the subject acquires first-person epistemic self-knowledge is supposed to be a further item different from the first-person avowal itself, neo-expressivist accounts endorse a *relational view of expression*, according to which mental states and their characteristic set of expressions are two different items related in some way. Therefore, neo-expressivist accounts, just as epistemic accounts of Transparency, endorse both an epistemic view of first-person self-knowledge and a relational view of expression.

The label "behavioural expressivism", used to name the version of expressivism that is going to be defended in this essay (which endorses a semantic view of Transparency, a non-relational view of expression and a non-epistemic view of first-person self-knowledge), is meant to differentiate that version of expressivism both from neo-expressivism and from logical behaviourism (e.g., Carnap, 1995; Hempel, 1980). On the one hand, behavioural expressivism is different from neo-expressivism because it considers that mental states are identical to patterns of expressive behaviour and because it considers that first-person self-knowledge is only a matter of expression and not a matter of having a true warranted belief (i.e., epistemic self-knowledge). On the other hand, behavioural expressivism is different from logical behaviourism because it considers that mental states are *sui generis* expressive processes that are not reducible to physical processes; in other words, it considers that patterns of expressive behaviour (i.e., mental states) are *sui generis* expressive processes, and so, that they are not reducible to patterns of physical behaviour.

Against logical behaviourism, *physical items* (i.e., physical objects, facts or events) are not the only class of items (i.e., objects, facts or events) that exists in the world. There are different *classes* of items in the world because there are different *classes* of causal networks in the world not reducible to each other. To give some examples, material items[2] (e.g., water, storms or neurophysiological states), social items[3] (e.g., patriarchy, nations or birth rates) and

[2] The label "material items" is meant to include all the items (i.e., objects, facts or events) characteristic of natural sciences, for instance, physical items (e.g., gravity), chemical items (e.g., the dissolution of salt in water), biological items (e.g., a neurophysiological state), and so on. I don't want to imply, though, that physical, chemical or biological items belong to the same class of items or that they take place in the same class of causal network.

[3] As in the latter case, the label "social items" is meant to include all the items (i.e., objects, facts or events) characteristic of social sciences (e.g., historical items, economical items, sociological items, etc.) without implying that they are the same class of items.

expressive items[4] (e.g., a smile of happiness, a traffic sign or a poem) are three different classes of items of the world because they are items that occupy a place in three different classes of causal networks irreducible to each other: the causal network of material items, the causal network of social items and the causal network of expressive items, respectively. Let's describe some examples in more detail. Firstly, the birth rate in Spain is of 1.49 children per woman. This is a *social fact*, different from a material fact (you won't see a woman of flesh and bones with 1.49 children), because it is a fact that occupies a particular place in the *causal network of social items*: among the causes of such a low birth rate might be the precarious labour conditions of younger generations in Spain (social fact) and among the effects of such a low birth rate might be the impossibility to pay the pensions of the older generations (social fact). Secondly, a certain distribution of the flesh and muscles a subject's face with the figure of a smile is a *material fact* because it occupies a particular place in the *causal network of material items*: it is caused by a nerve impulse (material event) and it might cause a certain reflection of the light that can be registered by the retina of an eye (material event). Thirdly, a smile of happiness is an *expressive fact* because it occupies a particular place in the *causal network of expressive items*: the cause of the smile of happiness might be that the subject saw a friend after a long time (expressive event) and it might cause that the person at whom the smile was directed to also smiles in response (expressive event).

Even if it is true that sometimes a single item can occupy a place in two or more classes of causal networks at once, these classes of causal networks are irreducible to each other because they describe quite different causal chains. For instance, a smile can be both a material item (i.e., a certain distribution of the flesh and muscles of a face) and an expressive item (i.e., a smile of happiness) because it can occupy a particular place both in the causal networks of material items (i.e., it is caused by a nerve impulse and it causes a certain reflection of the light) and in the causal network of expressive items (i.e., it is caused by seeing a friend and it causes another smile of happiness) at once. However, even if the same item (i.e., the smile) occupies a place in two classes of causal networks at once, these two classes of causal networks are irreducible to each other because they describe different chains of causes and effects. For instance, the nerve impulse that causes the smile as a material item doesn't play any role in the causal network of expressive items (even if it is an enabling condition) because it doesn't cause

[4] As in the latter cases, the label "expressive items" is meant to include all the items (i.e., objects, facts or events) characteristic of expression and humanities (e.g., arts, mental states, signs, etc) without implying that they are all the same class of items.

the smile as an expressive item (i.e., as a smile of happiness); the smile as an expressive item (i.e., as a smile of happiness) is caused by seeing a friend. Another example: the act of seeing a friend after a long time doesn't play the same causal role in the causal network of expressive items and in the causal network of material items. In the causal network of expressive items, the act of seeing a friend after a long time *directly* causes (i.e., without the mediation of any additional cause) the smile as an expressive item (i.e., as a smile of happiness); while in the causal network of material items, the act of seeing a friend after a long time causes the smile as a material item (i.e., as a certain distribution of the flesh and muscles of the face of a subject) only *indirectly* (i.e., with the mediation of multiple additional causes, such as the light hitting the retina of the eyes, nerve impulses, a certain treatment of the information by the nervous system, etc.).[5]

The aim of this essay is to enrich and to defend the behavioural-expressivist account of Transparency. On the one hand, it is going to be argued that the non-relational view of expression and the expressive view of first-person self-knowledge that follow from the behavioural-expressivist account of Transparency are better than the relational view of expression and the epistemic view of first-person self-knowledge endorsed by neo-expressivist accounts. On the other hand, it is going to be argued that the behavioural-expressivist account of Transparency is better than epistemic accounts of Transparency because, in addition to the fact that there are independent reasons to prefer a semantic account of Transparency over an epistemic account, the phenomena of self-deception and Moore's paradox are appropriately explained when the behavioural-expressivist view of Transparency is endorsed. In order to develop these arguments, this essay is going to have the following structure.

In the first chapter, the state of the art of the philosophical discussion about Transparency is going to be described, and so, both epistemic accounts of Transparency and the behavioural-expressivist account of Transparency are going to be explicated in detail. The description of the state of the art on Transparency will show that there are already two reasons to prefer the behavioural-expressivist account of Transparency over an epistemic account. On the one hand, epistemic accounts of Transparency seem to have problems explaining first-person self-knowledge of already held attitudes with the alleged first-person procedure of

[5] Notice that this ontology is very different from Cartesian dualism. According to Cartesian dualism, mental items and material items are different classes of items because they belong to two different and (almost) disconnected realities: the thinking substance and the extended substance. However, in the ontology described here, mental items, material items, social items… are different classes of items that belong to the same reality, as it is proved by the fact that a single item can belong to two or more different classes of causal networks at once.

epistemic self-knowledge that they consider to be responsible for Transparency (Cassam, 2004; Gertler, 2011). On the other hand, it seems that the behavioural-expressivist account of Transparency is able to offer a more plausible view of the first-person deliberation about whether p involved in the phenomenon of Transparency (García, 2019a) because it claims that first-person avowals of attitude are judgements about whether p rather than self-ascriptions of attitudes (as epistemic accounts of Transparency claim).

In the second chapter, the behavioural-expressivist account of expression and first-person self-knowledge that follows from the behavioural-expressivist account of Transparency are going to be explicated and compared with neo-expressivist accounts of expression and first-person self-knowledge. On the one hand, in the first part of the chapter, it is going to be argued that endorsing a non-relational view of expression (i.e., that a mental state and its set of expressions are one and the same item) or a relational view of expression (i.e., that a mental state and its set of expressions are two different items) depends on the way in which it is understood that expressions are *evidence* of mental states: from the idea that expressions are *symptoms* or *defeasible criteria* of mental states follows a relational view of expression and from the idea that expressions are *indefeasible criteria* of mental states follows a non-relational view of expression. Then, the relational view endorsed by neo-expressivism and the non-relational view endorsed by behavioural expressivism are going to be described and it is going to be argued that the non-relational view is better than the relational view because it is able to explain the phenomena of *pretence* (i.e., to feign that one has a mental state that one doesn't actually have) and *dissimulation* (i.e., to hide that one has a mental state that one actually has) with less theoretical resources than the relational view of expression.

On the other hand, in the second part of the chapter, the neo-expressivist account of first-person epistemic self-knowledge is going to be described in detail. Then, it is going to be argued that the idea of first-person epistemic self-knowledge is conceptually flawed, and so, a behavioural-expressivist alternative is going to be proposed. In line with behavioural expressivism, it is going to be argued that *first-person self-knowledge* is *expressive* self-knowledge (i.e., self-knowledge that has to do with the exercise of the ability to express our mental states) and that *third-person* self-knowledge is *epistemic* self-knowledge (i.e., true warranted belief about one's own mental states), which can be *authoritative* on those occasions in which the subject manages to warrant the belief about her own mental states to a higher degree than the beliefs of other people about her own mental states can possibly be.

In the third chapter, it is going to be argued that from the behavioural-expressivist account of Transparency follows the best account of self-deception among the accounts of self-deception currently available in the literature. Self-deception (e.g., Davidson, 2004; Fernández, 2013; Mele, 2001) is a motivated state characterized by an irrational conflict between what the subject sincerely says (e.g., "I believe that I am healthy") and how he acts (e.g., avoiding medical appointments or talks about medical issues, enhancing his health insurance more than he is able to pay, getting suddenly interested in the possibility of an afterlife, and so on) that reveals some kind of lack of self-knowledge. It has been argued (e.g., Gendler, 2010) that self-deception is a *sui generis* mental state; it has been argued as well that self-deception is a certain process that generates a false or unwarranted belief because of the intention of the subject to deceive himself (intentionalist accounts; e.g., Davidson, 2004), because of a motivated bias in the deliberation about whether p (motivationalist accounts; e.g., Mele, 2001) or because of an epistemic failure in the Transparency procedure responsible for first-person epistemic self-knowledge (epistemic accounts; Fernández, 2013). However, in line with the behavioural-expressivist account of Transparency, it is going to be argued in this chapter that self-deception is an unconscious mental state that cannot be self-consciously nor appropriately expressed. Then, it is going to be argued that self-deceivers cannot have first-person expressive self-knowledge of their states of self-deception because they can answer the transparent deliberative question "Do you believe that p?" only with a non-self-conscious judgement about whether p, and that self-deceivers have difficulties acquiring third-person epistemic self-knowledge of their states of self-deception because they have difficulties answering the non-transparent self-ascriptive question "Do you believe that p?" with a true self-ascription of attitude made on the basis of evidence about their own mental states (self-inspection).

And in the fourth chapter, it is going to be argued that from the behavioural-expressivist account of Transparency follows the best account of Moore's paradox (Moore, 1993) among the accounts of Moore's paradox currently available in the literature. Moore's paradox arises because sentences like "It is raining, but I don't believe so" and "It is raining, but I believe that it isn't" (*Moore's sentences*, henceforth) are irrational to assert even if they are sentences that have possible truth-conditions: it could be the case that it is raining but I don't have any belief about the issue or I believe that it is not raining. Most of the accounts of Moore's paradox currently available assume, like epistemic accounts of Transparency, that first-person avowals are self-ascriptions of mental states different from assertions, and they try to explain the paradox arguing that Moore's sentences are irrational to assert because they don't have appropriate assertion-conditions (pragmatic accounts; e.g., Rosenthal, 2005), because their assertion or judgement involves inconsistent conscious mental states or inconsistent commitments (psychological accounts; e.g., Coliva, 2016) or because their assertion or judgement involves a failure in the Transparency procedure responsible for first-person epistemic self-knowledge (epistemic accounts; e.g., Fernández, 2013). However, in line with the behavioural-expressivist account of Transparency, it is going to be argued in this chapter that a semantic account of Moore's paradox is able to explain the phenomenon in an appropriate way. On the one hand, when "I don't believe that it is raining" or "I believe that it is not raining" are answers to the transparent deliberative question "Do you believe that p?", they are judgements about whether p, and so, there is a contradiction or a contradiction-like with the other part of the Moore's sentence (i.e., "p"). As a result, the irrationality of asserting or judging a Moore's sentence arises. On the other hand, when "I don't believe that it is raining" or "I believe that it is not raining" are answers to the non-transparent self-inspective question "Do you believe that p?", they are self-ascriptions of belief, and so, there is no contradiction or contradiction-like with the other part of the Moore's sentence (i.e., "p"). As a result, no irrationality arises at asserting or judging a Moore's sentence, and it is explained why Moore's sentences have possible truth-conditions.

Let's start by explicating in detail the different accounts of Transparency and how the behavioural-expressivist account already has some explicative advantages over epistemic accounts.

1. Transparency of Belief

We sometimes answer the question “Do you believe that p?” as if we were answering the world-directed question “Is p the case?”. Instead of looking “inward” to find out whether I believe that p (as “Do you believe that p?” may suggest), we usually look “outward” to the world to find out whether p is the case (as “Is p the case?” suggests). Then, it is said that the question “Do you believe that p?” is sometimes *transparent* to the world-directed question “Is p the case?”, both in the sense that the former is answered by the same procedure as the latter (i.e., considering the epistemic grounds or reasons for and against the fact that p) and in the sense that they have to be answered equally (i.e., “Yes”, “No” or “I don’t know”). This phenomenon is called “Transparency of belief”, and it was popularised in the context of the discussion about self-knowledge by G. Evans[6] and his famous example of the third world war:

> “[I]n making a self-ascription of belief, one’s eyes are, so to speak, or occasionally literally, directed outward –upon the world. If someone asks me ‘Do you think there is going to be a third world war?’, I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question ‘Will there be a third world war?’ I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p. […] If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states.” (Evans, 1982, p. 225)

[6] Although it was first pointed out by Roy Edgley (1969).

Transparency of belief is an only first-person phenomenon, and so, it can be used to mark the distinction between the *deliberative first-person perspective* and the *self-inspective third-person perspective*. When a subject answers the question "Do you believe that p?" by deliberating about whether p (i.e., considering the grounds or reasons for and against the fact that p), Transparency takes place and the question is answered from the first-person perspective. So, the subject in Evan's example is supposed to have answered the question "Do you think there is going to be a third world war?" from this first-person deliberative perspective. However, when a subject answers the question "Do you believe that p?" by self-inspecting herself in order to make a judgement about whether she currently *believes* that p (i.e., on the basis of her own behaviour, opinions, inner-speeches, imaginings, feelings, passing thoughts or friends' testimonies), Transparency doesn't take place and the question is answered from the third-person perspective.[7] Certain examples of implicit bias are useful to show the differences between the deliberative first-person perspective and the self-inspective third-person perspective.

Imagine the following case of implicit bias. Tom is a person who allegedly believes in gender equality. When he deliberates about whether men are somehow superior to women, he finds no reason to think that they are. For instance, when he considers whether men are better drivers than women, he remembers that the statistics show that men have slightly more crashes than women; when he considers whether men are intellectually more capable than women, he remembers that there are slightly more women than men graduating from Spanish universities each year; and when he considers whether men have some kind of natural right to be fed by their female partners, he disregards that idea as stupid and crazy. True, Tom admits that there are currently more men than women in positions of power or big responsibility (e.g., as CEOs, as police officers, or as political leaders) and that there are currently more women than men working in caring professions (e.g., teachers, nurses, or babysitters), when not fully dedicated to domestic chores (e.g., housekeepers); but Tom understands those differences as effects of patriarchy (i.e., a cultural and symbolic system of domination, rooted in our societies, that creates different kinds of social roles for males and for females, with the roles granted to males being dominant over the roles granted to females) rather than as effects of natural or biological

[7] Notice that self-inspection is a form of deliberation. However, for the sake of clarity, I will reserve the term "deliberation" for when the question "Do you believe that p?" is answered from the first-person perspective (i.e., by deliberating about *whether p* is the case) and the term "self-inspection" for when the question "Do you believe that p?" is answered from the third-person perspective (i.e., by deliberating about whether *one believes* that p).

differences (in capabilities, in interests, in yearnings, etc.) between the two sexes. As a result, each time that Tom deliberates about gender equality, he firmly and sincerely ends up judging "[I believe that][8] men and women should be treated equally because they are equal".

However, things are different when Tom moves on from abstracts considerations about gender equality and has to deliberate about everyday situations in which both men and women happen to be involved. For instance, when Tom deliberates about which of his friends is a better driver, he tends to underestimate the capabilities of those friends that happen to be women (even if, as a matter of fact, all of them are good drivers); when Tom deliberates about who is the better candidate to fill an important vacancy at his company, he tends to underestimate the capabilities of those candidates that happen to be women (even if all candidates have a very similar CV); and when he considers whether his friends brought enough homemade food to share at the party, he tends to be more indulgent with male-friends than with female-friends (even if they all partake with a very similar contribution, both in terms of quality and quantity). So, Tom has an implicit bias in regard to gender. Even if he judges that men and women should be treated equally because they are equal, he is implicitly biased when he has to take part and to deliberate about issues that apparently have nothing to do with gender but in which both men and women are involved.

As it was said before, cases of implicit bias are useful to show why Transparency is an only first-person phenomenon. Imagine that Tom has to answer the question "Do you believe that men and women are equal?". Tom can answer this question both by deliberating about whether men and women are equal (first-person perspective) and by self-inspecting himself in order to make a judgement about what he believes about gender equality (third-person perspective). If Tom answers the question "Do you believe that men and women are equal?" from the first-person perspective, he will answer in a transparent way (i.e., as if he were answering the world-directed question "Is p the case?"), for he will have to deliberate about whether men and women are equal as a matter of fact. Since he won't find any good reason to doubt that men and women are equal, he will conclude the deliberation judging "[I believe that] men and women are equal". However, if Tom answers the question "Do you believe that men and women are equal?" from the third-person perspective, he will answer in a non-transparent way (i.e., in a different way than the world-directed question "Is p the case?"), for he will have

[8] Notice that, in this context, Tom can say to conclude the deliberation both "*I believe* that men and women should be treated equally because they are equal" and "Men and women should be treated equally because they are equal" without any change in the propositional content of the utterance.

to self-inspect himself in order to make a judgement about which attitude he holds towards gender equality on the basis of evidence about how he interacts with men and women. For instance, Tom could think about how nervous he gets when he has to get into a car driven by one of his female friends and about how that feeling could consistently draw his attention towards the possible dangers of the road, thus biasing his judgement about the expertise of female drivers; he could think about the suspicious fact that male job-candidates usually strike him as more capable than female job-candidates even after realizing, from careful scrutiny, that they all have a very similar CV; or he could think about how his female friends always complain that he is more demanding with them than with the men of the group when they appear at a party with their homemade food to share. As a result, Tom will conclude the self-inspection by making a judgement, not about whether men and women are equal, but about whether he actually *believes* that men and women are equal; e.g., "I don't actually believe that men and women are equal because people who actually believe so don't act as I do in those everyday situations". If this judgement is true and if it is made on the basis of good evidence, Tom will acquire *third-person self-knowledge* about the fact that he isn't an egalitarian person after all because he has an implicit bias against women, an implicit bias that a real egalitarian person wouldn't have (at least, not to that degree).

All the scholars on Transparency that I know of could agree on the description of the phenomenon that we've just seen (after making the desired terminological adjustments). However, there is controversy on the way to understand Transparency from this point on. On the one hand, following Evans, it is usually understood that Transparency is an *epistemic phenomenon* arising because of the special first-person procedure that human beings use to acquire *epistemic self-knowledge* of their own attitudes (e.g., Boyle, 2009, 2011, 2015; Byrne, 2005, 2011, 2018; Evans, 1982; Fernández, 2013; Gallois, 1996; Moran, 2001, 2003). These epistemic accounts of Transparency consider that the question "Do you believe that p?" asks about the *subject's beliefs* both when it is meant in a deliberative way and it is supposed to be answered with a first-person avowal, and when it is meant in a self-ascriptive way and it is supposed to be answered with a third-person avowal. However, epistemic accounts consider that Transparency occurs *only* when the question "Do you believe that p?" is answered from the first-person deliberative perspective because only under this condition is the subject supposed to apply the special first-person procedure of epistemic self-knowledge to answer the question "Do you believe that p?" by deliberating about whether p (i.e., as if she were answering the different question "Is p the case?").

On the other hand, against epistemic accounts of Transparency, the behavioural-expressivist account understands Transparency as a *semantic phenomenon* that has nothing to do with epistemic self-knowledge (García, 2019a). Transparency arises because the question "Do you believe that p?" has two different meanings depending on whether it is meant in a *deliberative* or in a *self-ascriptive* way. When the question "Do you believe that p?" is meant in a deliberative way, it is an invitation for the addressee to deliberate and to make up her mind about whether p, being so semantically equivalent to the question "Is p the case?". Since the question is meant to be answered by deliberating about whether p, it is meant to be answered from the first-person perspective; since the meaning of the question "Do you believe that p" is equivalent to the meaning of the question "Is p the case?", it is explained why Transparency arises here. However, when the question "Do you believe that p?" is meant in a self-ascriptive way, it is meant as an invitation for the addressee to tell us by self-inspection what she already believes about p (if anything), so the question "Do you believe that p?" semantically differs here from the question "Is p the case?". Since the question is meant to be answered by self-inspection, it is meant to be answered from the third-person perspective; since the question "Do you believe that p?" has a different meaning than the question "Is p the case?", it is explained why Transparency does not arise here.

The aim of this chapter is to describe the state of the art on Transparency and to argue that it shows that there are already reasons to prefer the behavioural-expressivist account of Transparency over epistemic accounts. The chapter is going to have the following structure. Firstly, the chapter is going to explicate the main types of epistemic accounts of Transparency, the main arguments in favour of them and an objection against them. Secondly, the chapter is going to explicate in detail the behavioural-expressivist account of Transparency, how it manages to refute the arguments in favour of epistemic accounts, and how it explains Transparency in a more plausible and intuitive way.

1.1 Epistemic accounts of Transparency

Epistemic accounts identify Transparency with a special first-person procedure that consists in a deliberation about whether p based on epistemic grounds or reasons and that delivers epistemic self-knowledge. Depending on how this first-person deliberation is characterized, two main types of epistemic accounts of Transparency can be identified: *spectatorial views* (Byrne, 2005, 2011, 2018; Fernández, 2013; Gallois, 1996) and *agential views* (Boyle, 2009, 2011, 2015; Moran, 2001, 2003). On the one hand, spectatorial views consider that Transparency is the deliberative first-person procedure by which subjects normally form second-order beliefs and that it can deliver first-person epistemic self-knowledge because of the special way in which it normally *warrants* the subject's second-order beliefs. On the other hand, agential views consider that Transparency is a procedure that can deliver first-person epistemic self-knowledge because, insofar as subjects are agents of their own attitudes, subjects have the *epistemic right*[9] to make judgements about their own attitudes by deliberating about their subject matter (e.g., whether p).

Among the accounts that defend a spectatorial view of Transparency are the *bypass account* (Fernández) and the *doxastic schema account* (Byrne and Gallois). Fernández claims that Transparency is a first-person procedure of belief-formation that consists in by-passing the *first-order belief that p* to form the *second-order belief that I believe that p* on the basis of the very *same grounds* on which one has formed the first-order belief that p. Let's see an example. Imagine that I look through the window and I see (or I seem to see) that it is raining so that I form the first-order belief that it is raining on the basis of that perceptual experience. The bypass procedure describes that, on the basis of the very same grounds on which I have formed the first-order belief that it is raining, I normally form the second-order belief that I believe that it is raining. Then, I normally form the second-order belief that I believe that it is raining on the basis of the very same perceptual experience of rain on which I have formed the first-order belief that it is raining. By contrast, Byrne and Gallois claim that Transparency is a procedure of belief-formation that consists in reasoning following the doxastic schema "If p, believe that you believe that p". According to this rule, each time that I judge that p is the case (forming the first-order belief that p), I am entitled to infer from that premise that I believe that p (forming the second-order belief that I believe that p). For instance, imagine that I see that it is sunny today and I form the first-order belief that it is sunny today by judging "It is sunny

[9] "What right have I to think that my reflection on the reasons in favor of p (which is one subject-matter) has anything to do with the question of what my actual belief about p is (which is quite a different subject-matter)? Without a reply to this challenge, I don't have any right to answer the question that asks what my belief is by reflection on the reasons in favor of an answer concerning the state of the weather. And then my thought at this point is: I would have a right to assume that my reflection on the reasons in favor of rain provided me with an answer to the question of what my belief about the rain is, if I could assume that what my belief here is was something determined by the conclusion of my reflection on those reasons." (Moran, 2003, p. 405).

today". In this situation, I am entitled (by the doxastic schema) to infer that *I believe* that it is sunny today from the premise that it is sunny today, acquiring so the second-order belief that I believe that it is sunny today.

Both the *bypass procedure* and the *doxastic schema procedure* are supposed to explain the phenomenon of Transparency and the phenomenon of first-person epistemic self-knowledge. On the one hand, they are supposed to explain Transparency because they describe that I normally form the second-order belief that I believe that p (which is supposed to answer the question "Do you believe that p?") by attending either to the grounds on which I have formed the first-order belief that p (bypass procedure) or by inference from my first-order belief that p itself (doxastic schema). As a result, if everything goes as it should, both procedures answer the question "Do you believe that p?" with a second-order belief that involves that the first-order belief that p (which is supposed to answer the question "Is p the case?") has already been formed. On the other hand, both the bypass procedure and the doxastic schema procedure are supposed to explain first-person epistemic self-knowledge as well because they are supposed to explain its groundless and authoritative character. Firstly, they explain that first-person epistemic self-knowledge is groundless because the bypass procedure and the doxastic schema procedure deliver true warranted second-order beliefs on the basis of grounds about whether p and not about the subject's beliefs. Secondly, they explain that first-person epistemic self-knowledge is authoritative because the warrant of the second-order belief delivered by the bypass procedure and the doxastic schema procedure is *stronger* than the warrant of the beliefs delivered by other epistemic procedures (e.g., perception or inference). The reason why the bypass procedure and the doxastic schema procedure are supposed to deliver *strongly* warranted second-order beliefs is that they are procedures of belief-formation that are *more reliable* and *less prone to error* (i.e., they deliver true beliefs more often) than other procedures of belief-formation (e.g., perception or inference). As a result, the bypass procedure and the doxastic schema procedure are supposed to be able to warrant my second-order beliefs about my own mental states better (i.e., in a stronger way) than other people's beliefs about my mental states can possibly be.

Let's see why the bypass procedure and the doxastic schema procedure are supposed to warrant the subject's second-order beliefs in a stronger way than other procedures of belief-formation. Starting with the bypass procedure, Fernández argues that there are plenty of epistemic errors that could affect the warrant and the truth-value of beliefs formed by perception or inference, but that cannot affect the warrant or the truth-value of the second-order beliefs formed by the bypass procedure. In regard to perception, imagine that I'm spending the evening in the countryside and, in poor lighting conditions, I look at a bush that accidentally has similar a shape and colours to those of a sheep. Imagine that, in spite of the fact that the lighting conditions are poor, I form by perception the first-order belief that there is a sheep in the middle of the field on the basis of that perceptual experience. In this case, my first-order belief that there is a sheep on the field is both *unwarranted* and *false*. It is unwarranted because I formed that belief on the basis of a perceptual experience taken in poor lighting conditions and it is false because there isn't any sheep on the field, but rather a bush with a similar shape and colour. However, if I apply the bypass procedure to form the second-order belief that I believe that there is a sheep on the field on the basis of that perceptual experience, my second-order belief will be warranted (i.e., I formed it on the basis of the very same grounds on which I formed the relevant first-order belief) and true (i.e., I have the first-order belief that there is a sheep on the field). In regard to inference, imagine that I infer on the basis of a fortune-teller (i.e., poor epistemic grounds) that it is going to rain tomorrow. My first-order belief that it is going to rain tomorrow is unwarranted and it can be false if it happens that it doesn't rain tomorrow. However, if I apply the bypass procedure to form the second-order belief that I believe that it is going to rain tomorrow on the basis of what the fortune-teller told me, my second-order belief will be warranted (i.e., I formed it on the basis of the very same grounds as the relevant first-order belief) and true (i.e., I actually have the first-order belief that it is going to rain tomorrow). Therefore, since neither the warrant nor the truth of second-order beliefs are affected by typical epistemic errors at first-order level, the bypass procedure is more reliable and less prone to error than perception or inference (methods that other subjects can use to forms beliefs about my own mental states), and so, it is explained why the bypass procedure can deliver second-order beliefs warranted in a stronger way than perception or inference.

Secondly, Byrne argues that the doxastic schema (i.e., "If p, believe that you believe that p") is a highly reliable method of belief-formation because it is *self-verifying*: when it is properly applied by the subject, it cannot deliver a false second-order belief. Indeed, even if the doxastic schema itself is not warranted either deductively (i.e., it is clear that p can be the case without being the case that I believe that p) or inductively (i.e., there are plenty of facts about which I don't have any belief), the doxastic schema is self-verifying because its correct application guarantees that the subject ends up both with a second-order belief and with the particular first-order belief that is the truth-maker of that second-order belief. For the doxastic schema claims that if I judge that p, I have to infer the second-order belief that I believe that p; and if I judge that p, I already have the first-order belief that p (assuming that judging that p involves believing that p), which is the truth-maker of the second-order belief that I believe that p. As a result, the doxastic schema can deliver false second-order beliefs only under the condition that the subject makes a mistake in the application of the process, being so immune to epistemic failures caused by a lack of "collaboration" of the object of knowledge and its misleading properties (i.e., "brute errors"). Indeed, in cases of perceptual or inferential knowledge, subjects can end up forming a false belief both because of a *mistake of the subject* (e.g., to form a false belief about the colour of a wall right after perceiving it in poor lighting conditions) and because of *the properties of the object* of knowledge (e.g., to form the false belief that there's a sheep on the field on the basis of a bush perceived in good lighting conditions but that happens to have the same shape and colour of a sheep). However, this latter kind of epistemic failures (brute errors) are ruled out in the application of the doxastic schema procedure because it is self-verifying when it is properly applied; the only possibility of epistemic failure here is that the subject fails at following the doxastic schema in an appropriate way. Thus, the true second-order beliefs delivered by the doxastic schema are warranted in a stronger way than perceptual or inferential beliefs because of their *safety*, meaning that they "could not easily have been false" (Byrne, 2018, p.110), and so, they are instances of first-person epistemic self-knowledge.

Therefore, both the bypass procedure and the doxastic schema procedure are supposed to explain first-person epistemic self-knowledge because they can deliver *strongly warranted true second-order beliefs* on the basis of no specific grounds about the subject's mental states. On the one hand, insofar as those second-order beliefs are not formed on the basis of grounds about the subject's mental states, it is explained why first-person epistemic self-knowledge is groundless. On the other hand, insofar as the bypass and the doxastic schema procedures are highly reliable processes of belief-formation (e.g., more reliable than perception or inference), it is explained why first-person epistemic self-knowledge is authoritative or presumably true. It is true that these second-order beliefs can still end up being false if the subject makes an epistemic error in the application of the bypass or the doxastic schema procedures, but when no epistemic error occurs, the bypass and doxastic schema procedures deliver first-person epistemic self-knowledge.

Moving on from spectatorial views, it is now the turn of agential views of Transparency (Moran, Boyle). Agential views consider that Transparency and first-person epistemic self-knowledge are explained by the *constitutive link* between our attitudes and our reasons to hold those attitudes. Subjects are sometimes victims of *alienated attitudes*, attitudes that by their own nature are not sensitive to their deliberative reasons, and so, attitudes that can't be self-known from the first-person perspective. Tom's case of implicit bias could be considered an example of alienated belief by agential views of Transparency insofar as it might be argued that Tom believes that men and women are not equal and that this belief is an alienated belief because it is not sensitive to Tom's deliberative reasons (Tom doesn't have reasons to believe, all things considered, that men and women are not equal). However, when subjects are responsible *agents* of the attitudes that they form and hold because such attitudes are sensitive to their deliberative reasons (i.e., they are not alienated), subjects have the epistemic right to make judgements about their own attitudes (e.g., about whether I believe that it is raining) on the basis of their deliberative reasons about the subject matter of those attitudes (e.g., on the basis of my reasons for and against the fact that it is raining) because there is a constitutive link between those deliberative reasons and the attitude held by the subjects (e.g., between my reasons for and against the rain and my belief that it is raining). So, in these cases, subjects can enjoy first-person epistemic self-knowledge of their attitudes.

Moran argues that both a *theoretical* and a *deliberative* stance can be adopted towards one's attitudes. When we adopt a theoretical stance towards one's attitudes, we consider our attitudes as autonomous entities disconnected from the deliberative reasons that we might have for holding them; i.e., we consider them by the effects that they have on us, such as a particular behaviour or a feeling about something. This is the stance characteristic of the third-person perspective, from which —according to Moran— we could acquire non-authoritative third-person self-knowledge if we discovered something true about ourselves. Among the mental states that we can discover from this theoretical point of view are the alienated attitudes that we mentioned above; attitudes that are not sensitive to the subject's deliberative reasons, and so, that are neither transparent to the world nor self-known from the first-person perspective. By contrast, when we adopt a deliberative stance towards one's attitudes, we consider our attitudes as transparent to the world because we consider them in connection with the deliberative reasons by which they are held. This is the characteristic stance of the first-person perspective, from which subjects can avow their mental states with first-person epistemic self-knowledge; among them, non-alienated beliefs. Indeed, non-alienated beliefs are sensitive to reasons by their own nature: they have been formed, and they continue to be held, by assessing the reasons for and against their subject matter (e.g., whether p is the case) until one's mind is made up about the issue.

The phenomenon of Transparency is readily explained from the deliberative stance. For in order to answer the question "Do you believe that p?" from the first-person deliberative stance, one has to follow the same procedure as the one needed in order to answer the question "Is p the case?": deliberate about whether p on the basis of reasons for and against until one forms a belief about p. Also, first-person epistemic self-knowledge is explained as well by the deliberative stance. For due to the constitutive link between beliefs and reasons, subjects can be aware of the beliefs that they form and hold through their reasons to hold them; they have the epistemic right to make judgements about their own mental states through the conclusions of their deliberations about whether p.

Boyle gives a detailed account of how the characteristic awareness of first-person epistemic self-knowledge must be understood. According to Boyle, reason-sensitive attitudes, and only reason-sensitive attitudes, are *self-reflective*. Self-reflective or self-aware attitudes (as opposed to alienated attitudes) are attitudes that, by their own deliberative nature, can be the *object of the subject's shifting of attention between the world and her mind*. Indeed, in Boyle's account, when I have a self-reflective attitude formed by deliberation on the basis of reasons (e.g., the self-reflective belief that it is raining), I am able to shift my attention from the aspect of the world to which I am engaged (e.g., that it is raining as a matter of fact) to my engagement or attitude itself (e.g., to my belief that it is raining). Thus, a *self-reflective* attitude is a single cognitive state with two different aspects: a judgement about the world made on the basis of deliberative reasons and the attitude or engagement to the world resulting from that judgement. That's why self-reflective attitudes, unlike alienated attitudes, have the capability of being the object of the subject's shifting of attention between the judgement about the world on the basis of which they are formed and the attitude or engagement to the world resulting from that judgement. In Boyle's words:

> "The reflective approach explains doxastic transparency […] as a matter […] of shifting one's attention from the world with which one is engaged to one's engagement with it – an engagement of which one was already tacitly cognizant even when one's attention was "directed outward." (Boyle, 2001, p. 228).

And,

> "On his view, the important truth is this: the very same actualization of my cognitive powers that is my believing P is, under another aspect, my tacitly knowing that I believe P. Hence, to pass from believing P to judging I believe P, all I need to do is reflect – i.e., attend to and articulate what I already know. Something broadly similar will hold for other psychological conditions of which I can have transparent self-knowledge." (Boyle, 2001, p. 229).

When I focus my attention on my attitude or engagement to the world, I have *actual awareness* of my self-reflective attitude because I am aware of the engagement that constitutes my attitude in this very same moment (e.g., I have actual awareness of my self-reflective belief that it is raining when I attend to my engagement to the fact that it is raining). However, when I focus my attention on the aspect of the world to which I am engaged (e.g., that it is raining), I have *tacit awareness* of my self-reflective attitude (e.g., the belief that p) in the sense that I am not aware of my attitude or engagement in this very same moment (even if I could shift my attention at any time to gain actual awareness again).

Therefore, agential views explain first-person epistemic self-knowledge of attitudes because they explain why first-person avowals are groundless and authoritative self-ascriptions of attitude. On the one hand, first-person avowals are groundless self-ascriptions of attitude because they are made on the basis of reasons about the subject matter of the self-ascribed attitude itself (e.g., about whether it is raining) and not on the basis of reasons about the subject's mental states (e.g., about whether I believe that it is raining). On the other hand, first-person avowals are authoritative or presumably true because only the subject who issue the first-person avowal can be aware of the self-reflective attitude that she holds through the deliberative reasons to hold it (because the constitutive link takes place only between *her* attitude and *her* deliberative reasons), and so, first-person avowals are warranted in a stronger way than third-person ascriptions of attitudes (i.e., they are more likely true).

To conclude, let us point out three fundamental ideas that can be tracked in every epistemic account of Transparency. The first idea is that the question "Do you believe that p?" asks about the *subject's beliefs* both in first-person deliberative conversational contexts in which Transparency arises (i.e., the question “Do you believe that p?” is answered in the same way as the question “Is p the case?”) and in third-person self-ascriptive conversational contexts in which Transparency doesn’t arise (i.e., the question “Do you believe that p?” is not answered in the same way as the question “Is p the case?” because it is answered by self-inspection). The second idea is that the answer to the question “Do you believe that p?” is a *self-ascription of attitude* both in first-person deliberative conversational contexts in which the answer to the question “Do you believe that p?” is a first-person avowal (so that Transparency arises and the question is answered by deliberating about whether p) and in third-person self-ascriptive conversational contexts in which the answer to the question “Do you believe that p?” is a third-person avowal (so that Transparency doesn’t arise because the question is answered by self-inspection). And the third idea is that first-person epistemic self-knowledge is responsible for filling the semantic gap existing between the transparent question “Do you believe that p?” and “Is p the case?” by allowing subjects to make groundless and authoritative self-ascriptions of mental states (i.e., first-person avowals) on the basis of epistemic grounds or reasons about whether p (i.e., as if they were answering the question “Is p the case?”). Therefore, Transparency occurs only in first-person deliberative contexts because only in first-person deliberative contexts the special procedure responsible for *first-person epistemic self-knowledge* can be applied.

In the following two sections, the main arguments in favour and against epistemic accounts of Transparency are going to be discussed.

1.2 Two arguments in favour of epistemic accounts

Two main arguments can be found in the literature about Transparency and first-person self-knowledge to make the case in favour of epistemic accounts. The first argument is developed by Fernández (2013, pp. 50-51) and it focuses on the meaning that the *question* “Do you believe that p?” has in different contexts. Fernández argues that there are conversational contexts in which the question “Do you believe that p?” is asked and two things occur at the same time. Firstly, Transparency occurs because the addressee answers the question “Do you believe that p?” from the first-person deliberative perspective (i.e., on the basis of grounds or reasons for and against the fact that p). Secondly, it is clear that the question “Do you believe that p?” is meant as a question about whether the addressee *believes that p* and not as a question about *whether p* is the case (in this way having a different meaning than the question “Is p the case?”). Fernández gives two examples of such type of conversational contexts:

> “[…] imagine a lawyer whose client claims to be innocent. It is important to the client that his lawyer believes him, so he asks his lawyer ‘Do you believe that I am innocent?’ Clearly, he is not asking whether he is innocent. He must already know that. […] Yet, if the lawyer has not reflected on her attitudes towards her client before, then what she will do to address the question is to focus on those considerations that would support or challenge the belief that her client is innocent. Similarly, imagine that my priest asks me ‘Do you believe that God exists?’ Clearly, he is not asking me whether God exists. Presumably, his mind is already made up on that issue. […]. Yet, if I have never thought about whether I am a religious person or not, I will address his question by attending to the evidence for and against the existence of God.” (Fernández, 2013, p. 51)

Fernández argues that the questions “Do you believe that I am innocent?” and “Do you believe that God exists?” are supposed to be answered from a deliberative first-person perspective; i.e., on the basis of grounds or reasons as to whether the client is innocent or as to whether God exists. However, even if these questions are supposed to be answered in a transparent way, it is clear that they ask about the addressees’ beliefs (i.e., whether the lawyer *believes* that the client is innocent and whether the parishioner *believes* that God exists) and not about any other fact of the world (e.g., whether the client is innocent as a matter of fact or whether God exists as a matter of fact). For it is obvious for all the subjects involved in the conversation that the questioners already have made up their minds about whether the client (i.e., herself) is innocent as a matter of fact or about whether God exists as a matter of fact, and so, it is obvious for all the subjects involved in the conversation that what the questioners want to know is the *belief of the addressees* about those subject matters (i.e., one’s innocence or the existence of God) and not any other fact of the world. Therefore, we have here two clear cases in which 1) Transparency takes place and 2) the question “Do you believe that p?” ask about the subject’s beliefs and not about whether p.

To clarify, the point made here is not that *there aren't* conversational contexts in which we use the question "Do you believe that p?" to ask about the fact that p (e.g., imagine a context in which I ask "Do you believe it is raining?" being solely interested in knowing whether it is raining —indirect speech act—). Rather, the point made here is that *there are some* conversational contexts in which Transparency occurs and, as it is required by epistemic accounts, the question "Do you believe that p?" asks about the addressee's belief that p. So, Fernández concludes, the semantic gap between the questions "Do you believe that p?" and "Is p the case?" (which constitutes the starting point of epistemic accounts) turns out to be true. These questions have different meanings (i.e., they ask about different things), even in conversational contexts in which Transparency occurs, because the former asks about the addressee's belief that p and the latter asks about whether p is the case.

The second argument in favour of epistemic accounts of Transparency focuses on the meanings of the *answers* to the question "Do you believe that p?" in deliberative first-person conversational contexts. This argument is based in Bar-on's argument (Bar-on, 2004, pp. 8-9) in favour of the *semantic continuity* between first-person avowals (e.g., "I believe that p") and typical ascriptions of mental states made by third-person subjects about oneself (e.g., "Jesús believes that p"). For, if first-person avowals are self-ascriptions of mental states just as ascriptions of mental states made by third-person subjects (semantic continuity), it follows the idea (endorsed by epistemic accounts of Transparency) that first-person avowals, which answer the transparent question "Do you believe that p?", are self-ascriptions of mental states rather than judgements about whether p.

Bar-on's *semantic continuity* argument goes as follows. There are two reasons why it is clear that there is a semantic continuity between first-person avowals and ascriptions of mental states made by third-person subjects. On the one hand, the first reason is that both first-person avowals and ascriptions of mental states made by third-person subjects can be exchanged *salva veritate* in similar contexts, and hence, both kinds of utterances must have the same truth-conditions and the same semantic content. For instance, if my first-person avowal "I believe that p" is true, it follows that someone can truly say about me "He believes that p" or "Jesús believes that p". On the other hand, the second reason why first-person avowals and ascriptions of mental states made by third-person subjects must have the same semantic content is that subjects can use any first-person avowal of mental state as a premise in sound arguments. For instance, I can make an argument such as the following: 1) "If I believe that it is raining, I should advise my flatmate to pick up the umbrella"; 2) "I believe that it is raining"; 3) "Hence, I should advise my flatmate to pick up the umbrella”. If this argument is sound, and it actually is, the truth of the premise “I believe that it is raining” (which is supposed to be a first-person avowal) has to be transmitted to the conclusion. Moreover, since the argument only works when *I believe* that it is raining (for it is not expected that I can warn my flatmate when it is raining but I don’t believe so), the truth of the premise “I believe that it is raining” (which is supposed to be a first-person avowal) has to do with *my belief* that p and not with any other fact of the world. Therefore, it is concluded, first-person avowals are *self-ascriptions* of mental states and their meanings are about the subject’s mental states and not about any other fact of the world.

From this argument follows that there is a semantic gap between first-person avowals (e.g., “I believe that p”) and assertions (e.g., “p is the case”). This semantic gap, in turn, justifies the semantic gap between the transparent question “Do you believe that p?” (which is allegedly answered by a first-person avowal) and the question “Is p the case?” (which is allegedly answered by an assertion). Therefore, the semantic continuity argument can be used to vindicate the starting point of epistemic accounts of Transparency (even if that wasn’t Bar-on’s original intention).

1.3 One argument against epistemic accounts

It has been argued (Cassam, 2004; Gertler, 2011) that epistemic accounts of Transparency cannot explain how we acquire first-person epistemic self-knowledge of all the attitudes that we are supposed to have first-person epistemic self-knowledge of. Granting — for the sake of argument— that epistemic accounts could explain how we acquire first-person epistemic self-knowledge of our attitudes *at the moment in which they are formed*, it seems that they cannot explain how we could have first-person epistemic self-knowledge of the attitudes that we *already hold* because we have formed them in the past. This objection is relevant against epistemic accounts of Transparency because epistemic accounts think that first-person epistemic self-knowledge is a genuine phenomenon and that an appropriate account of Transparency has to explain first-person epistemic self-knowledge as well (or, at least, first-person epistemic self-knowledge of attitudes). Let's see the objection using belief as an example.

Epistemic accounts identify Transparency with a first-person procedure responsible for epistemic self-knowledge (bypass, doxastic schema or agential deliberation). Then, when a subject follows the Transparency procedure, she is supposed to be able to acquire first-person epistemic self-knowledge of the belief that p on the basis of grounds or reasons for and against the fact that p. However, intuitively, there is a difference between having first-person epistemic self-knowledge of a new belief at the moment in which it has been formed and having first-person epistemic self-knowledge of an already held belief. For instance, if you ask me which political party I think is going to win the next elections and I've never thought about that before, I have to apply the Transparency procedure to form a belief about that matter by considering the actual chances that the different political parties have to win the elections. Epistemic accounts could explain how we are supposed to have first-person epistemic self-knowledge in this type of cases insofar as they consider that we can acquire first-person epistemic self-knowledge by deliberating about whether p (i.e., following Transparency). However, if you ask me which political party I think is going to win the next elections and I already have a belief about that matter because I already followed the Transparency procedure in the past, I am not supposed to apply the Transparency procedure *again* (i.e., I am not supposed to assess again the chances that the different political parties have to win the elections) to be able to have first-person epistemic self-knowledge of my belief and answer your question. Thus, once that one already believes that p, it doesn't seem necessary to consider again the grounds or reasons about whether p in order to have first-person epistemic self-knowledge and answer the question "Do you believe that p?".

Furthermore, it might be counterproductive to apply the Transparency procedure *again* to answer the question "Do you believe that p?" if what the questioner wants to know is what I *already* believe. For by applying the Transparency procedure again (i.e., by assessing the chances again that the different political parties have to win the elections), I might end up forming a new belief instead of acquiring first-person epistemic self-knowledge of the belief that I already have. Thus, epistemic accounts of Transparency have troubles explaining how subjects seem to have first-person epistemic self-knowledge of *already held* beliefs.

Epistemic accounts may try to undermine this objection following some ideas about Transparency developed by Shah and Velleman (2005). Shah and Velleman argue that the question “Do you believe that p?” can mean either “Do you believe that p *right now*?” or “Do you *already* believe that p?”. We have seen that epistemic accounts could explain how subjects are supposed to have first-person epistemic self-knowledge of whether they believe that p *right now*. Indeed, since “right now” doesn’t discriminate whether the belief was already held or whether it is a new belief that has just been formed deliberating about whether p, subjects are supposed to be able to acquire first-person epistemic self-knowledge when they answer the question “Do you believe that p right now?” by applying the Transparency procedure. The problem arises, however, with the second question “Do you *already* believe that p?”. Following Shah and Velleman, epistemic accounts could argue that subjects can also acquire first-person epistemic self-knowledge by answering this second question in a transparent way. The only qualification is that subjects have to use a *slightly modified* Transparency procedure to answer the question “Do you already believe that p?” if they want to acquire first-person epistemic self-knowledge of whether they already believe that p. Indeed, in order to answer this question, subjects don’t have to consider again the grounds or reasons about whether p (in fact, this procedure could modify the belief that they already hold about p instead of enabling first-person epistemic self-knowledge of such belief); by contrast, they have to follow a slightly modified Transparency procedure in which answering the question “Do you already believe that p?” involves answering the question “Is p the case?” only insofar as subjects have to use the latter question as

> “[…] a stimulus applied to oneself for the empirical purpose of eliciting a response. One comes to know what one already thinks by seeing what one says—that is, what one says in response to the question *whether p*” (Shah and Velleman, 2005, p. 506).

Thus, in order to answer the questions “Do you believe that p right now?” and “Do you already believe that p?” from the first-person perspective and with first-person epistemic self-knowledge, subjects are supposed to follow some version of the Transparency procedure insofar as, in both cases, they have to answer the question “Is p the case?”. On the one hand, in order to answer the question “Do you believe that p right now?”, subjects have to deliberate about whether p as if they were answering the question “Is p the case?”. On the other hand, in order to answer the question “Do you already believe that p?”, subjects have to use the question “Is p the case?” as a stimulus to see what one spontaneously answers, only to find out on that basis what one already believes about the fact that p. Therefore, it can be concluded that epistemic accounts of Transparency could explain first-person epistemic self-knowledge of both new and already held beliefs.

However, this replica doesn't seem to be able to save epistemic accounts from the objection raised above. Once epistemic accounts have opened a semantic gap between the transparent question "Do you believe that p?" and the question "Is p the case?", they are forced to posit a procedure to acquire *first-person epistemic self-knowledge* that explains why subjects answer both questions in the same way only when they answer from the *first-person perspective*. And "to see what one says in response to the question *whether p*" doesn't seem like a good procedure to acquire *first-person epistemic self-knowledge* of the beliefs that one already holds but a good procedure to acquire *third-person epistemic self-knowledge* of the beliefs that one already holds. For "to see what one says in response to the question *whether p*" is a procedure that I can apply to myself just as much as other subjects can apply that procedure to me. So, "to see what one says in response to the question *whether p*" cannot be an epistemic procedure to acquire *first-person epistemic self-knowledge*, and this is exactly what epistemic accounts of Transparency need it to be in order to explain why Transparency is the *first-person* procedure responsible for epistemic self-knowledge. Therefore, no good explanation of how subjects are supposed to have *first-person* epistemic self-knowledge of their already held beliefs using the Transparency procedure has been given so far by epistemic accounts or their defenders.

In the following two sections, the behavioural-expressivist account of Transparency is going to be explicated in detail and it is going to be argued that the behavioural-expressivist account can explain the deliberative process behind Transparency better than epistemic accounts and that the behavioural-expressivist account can refute the arguments in favour of epistemic accounts while explaining the intuitions behind them.

1.4 The behavioural-expressivist account of Transparency

The behavioural-expressivist account (García, 2019a) considers that Transparency is a semantic phenomenon that arises because the question "Do you believe that p?" has different meanings in different types of conversational contexts. When “Do you believe that p?” is meant in a *deliberative way*, the meaning of that question is tantamount to the meaning of the question “Is p the case?”, and so, Transparency occurs: both questions are answered in the same way and by the same procedure because both questions ask about whether p. However, when the question “Do you believe that p” is meant in a *self-ascriptive way*, it is a question about the subject’s belief that p so that its meaning is different from the meaning of the question “Is p the case?”; hence, Transparency doesn’t occur: both questions are answered by different procedures because they ask about different things.

When the question “Do you believe that p?” is meant in a deliberative way, Transparency occurs because the question “Do you believe that p?” is meant as an invitation to deliberate and to make up one’s mind about whether p (i.e., just as the question “Is p the case?”). If the addressee understands that the question “Do you believe that p?” is meant in a deliberative way and if she follows through, two situations can take place: that she hasn’t made up her mind yet about the issue or that she has already made up her mind about the issue. On the one hand, if the addressee doesn’t have her mind already made up about the issue, she will deliberate about *whether p* on the basis of reasons for and against p and she will conclude that deliberation with a *judgement about whether p*. This judgement about whether p creates a new attitude about whether p (typically: belief, disbelief or suspension), and so, it is an *expressive episode* of the attitude that it creates as the conclusion of the deliberation (i.e., the first expressive episode of the *expressive pattern* of the newly formed attitude). On the other hand, if the addressee of the deliberative question “Do you believe that p?” has already made up her mind about the issue, she can do two things to answer the question: to start another deliberation about whether p from anew on the basis of reasons for and against, or to answer directly by expressing her current attitude about whether p on the basis of no extra reasons about whether p. If she decides to start another deliberation about whether p from anew, she will drop the already held attitude about whether p (typically: belief, disbelief or suspension) just to form a new attitude (typically: belief, disbelief or suspension) as the conclusion of the new deliberation. By contrast, if she decides to answer the question with the belief that she already holds, she will answer directly (i.e., without extra deliberation and on the basis of no extra reasons about whether p) with an expressive episode of the attitude about whether p that she holds since the moment that she formed it at the conclusion of her first-person deliberation in the past: with the same judgement about whether p by which she put a conclusion to the deliberation that gave rise to her current attitude about whether p. Of course, she can use reasons to defend her judgement about whether p, but only reasons that were already used in the original deliberation to support that judgement (otherwise, she would drop her current attitude and she would start another deliberation anew).

Anyway, the judgement about whether p by which the subject answers the question "Do you believe that p?" can take either the linguistic form of an avowal (typically: "I believe that p", "I believe that not-p" or "I don't believe either way") or the linguistic form of an assertion (typically: "p is the case", "p is not the case" or "It could be either way"). However, regardless of the linguistic form that it takes, this judgement about whether p is always an expressive episode of the expressive pattern characteristic of the attitude formed at the end of the deliberation about whether p (typically: belief, disbelief or suspension).

Let's explain how a subject can answer the question "Do you believe that p?" from the first-person deliberative perspective with an example. Imagine that someone asks me "Do you believe that the Earth goes around the Sun?" meant in a deliberative way and that I answer from the first-person deliberative perspective. Two situations can take place: either I don't have any belief about the issue (I have never thought about that or I made up my mind years ago and I don't remember anymore) or I already have an attitude about the issue. If I don't have any attitude about whether the Earth goes around the Sun, I will deliberate on the basis of reasons for and against whether the Earth goes around the Sun (e.g., that I read once that it does, that it seems to me that it is the Sun the one which moves, that I can reinterpret that visual appearance so that it is the Earth the one which moves, etc.) and I will conclude that deliberation with a judgement about whether the Earth goes around the Sun; judgement (e.g., the first-person avowal "I believe that the Earth…", the assertion "The Earth…", etc.) that is an expressive episode of the pattern of my newly formed attitude (typically: belief, disbelief or suspension). By contrast, if I already hold an attitude about whether the Earth goes around the Sun because I have already deliberated about the issue before and I remember the conclusion, I don't need to deliberate again on the basis of reasons for and against (what in fact would erase my current attitude) because I can answer the question directly (i.e., on the basis of no extra reason about whether p) with an expressive episode of the pattern of the attitude (typically, belief, disbelief or suspension) that I currently hold since I formed it at the conclusion of my first-person deliberation: with the same judgement about whether the Earth goes around the Sun by which I concluded the deliberation about whether the Earth goes around the Sun in the past (e.g., the first-person avowal "I believe that the Earth…", the assertion "The Earth…", etc.).

When the question “Do you believe that p?” is meant in a self-ascriptive way, it is meant as an invitation to answer with a *self-ascription of attitude*. If the addressee understands that the question is meant in a self-ascriptive way and if she follows through, two situations can take place: that she doesn’t have her mind already made up about the attitude that she holds (because she didn’t self-inspect herself about the issue before or because she self-inspected herself about the issue but she forgot) or that she has her mind already made up about the attitude that she holds (because she self-inspected herself about the issue in the past and she remembers). On the one hand, if the addressee hasn’t made up her mind yet about the attitude that she holds, she will self-inspect herself from the third-person perspective on the basis of evidence about her own mental states (e.g., her own feeling, actions, demeanour, passing thoughts, imaginings or utterances) and she will conclude this process of self-inspection with a judgement about which is the *attitude* that she currently holds about p (i.e., with a self-ascription of attitude). This judgement about the attitude that she holds creates a new *second-order belief*, and so, it is an *expressive episode* of the *pattern* of the second-order belief that it creates as the conclusion of the third-person process of self-inspection. On the other hand, if the addressee has already made up her mind about the attitude that she holds, she can do two things to answer the self-inspective question. Firstly, she can self-inspect herself from the third-person perspective again on the basis of evidence about her own mental states to form a new second-order belief about the attitude that she holds (in which case she will drop her current second-order belief just to form a new one at the end of the new self-inspective process), answering so with a new self-ascription of attitude (i.e., an expressive episode of the pattern of the new second-order belief). Secondly, she can answer directly (i.e., without self-inspecting herself again and on the basis of no extra evidence) with the same self-ascription of attitude or judgement about her own attitudes by which she concluded her third-person process of self-inspection in the past giving rise to the second-order belief that she currently holds (i.e., with an expressive episode of the pattern of her current second-order belief).

Anyway, the addressee’s self-ascription or judgement about the attitude that she holds is an expressive episode of the pattern of her second-order belief formed by self-inspection and it can take either the linguistic form of an avowal (e.g., “I believe that p”, “I believe that not-p” or “I don’t believe either way”) or the linguistic form of an assertion (e.g., “It is the case that I believe that p”, “It is the case that I believe that not-p” or “It is the case that I don’t believe either way”). However, regardless of whether it takes the form of an avowal or the form of an assertion, it is always a self-ascription of attitude, as it is revealed by the circumstances in which the utterance is made (i.e., as an answer to a self-inspective question about one's own beliefs) and by the circumstances in which the second-order belief is formed (i.e., as a conclusion of a third-person self-inspective process based on evidence about one's own mental states).

Tom's case of implicit bias is a clear example of how the question "Do you believe that p?" (i.e., "Do you believe in gender equality?") can be answered in different ways (e.g., "Yes", "No" or "I don't know") depending on whether it is answered from the first-person deliberative perspective (e.g., "Men and women are equal") or from the third-person self-inspective perspective (e.g., "I don't actually believe that men and women are equal"). However, it is not always the case that subjects have to give different responses (e.g., "Yes", "No" or "I don't know") to the question "Do you believe that p?" depending on whether they answer from the first-person deliberative perspective or from the third-person self-inspective perspective. To show this, let's continue with the example of the Earth and the Sun. Imagine that I answer the question "Do you believe that the Earth goes around the Sun?" from the third-person self-inspective perspective. So, instead of deliberating about whether the Earth goes around the Sun on the basis of reasons for and against (first-person deliberative perspective), or instead of answering directly (i.e., on the basis of no extra reason) with an expressive episode of the pattern of the attitude about whether the Earth goes around the Sun that I formed from a deliberation in the past (first-person deliberative perspective), I *judge that I believe* that the Earth goes around the Sun on the basis of the fact (evidence) that I remember myself explaining to my cousin how it is that the Earth goes around the Sun even if it seems otherwise from Earth's perspective. As a result, I answer the question "Do you believe that the Earth goes around the Sun?" from self-inspection with an expressive episode of the pattern of the newly formed second-order belief that I believe that the Earth goes around the Sun. This expressive episode consists in a self-ascription of belief, regardless of whether it takes the linguistic form of an avowal ("Yes, I believe that the Earth goes around the Sun") or an assertion ("It is the case that I believe that the Earth goes around the Sun").

Therefore, the behavioural-expressivist account explains Transparency attending to the semantic differences between the question "Do you believe that p?" when it is meant in a deliberative way and when it is meant in a self-ascriptive way. On the one hand, when the question "Do you believe that p?" is meant in a deliberative way, Transparency occurs (i.e., the question "Do you believe that p?" is answered in the same way and by the same procedure as the question "Is p the case?") because it is meant as a question about whether p, and so, it has the same meaning as the question "Is p the case?". On the other hand, when the question "Do you believe that p?" is meant in a self-ascriptive way, Transparency doesn't occur (i.e., the question "Do you believe that p?" is not answered in the same way as the question "Is p the case?") because it is meant as a question about the subject's beliefs, and so, it doesn't have the same meaning as the question "Is p the case?".

To conclude, let us point out the main differences between the behavioural-expressivist account of Transparency and epistemic accounts. Firstly, the behavioural-expressivist account claims that the question "Do you believe that p?" asks about *whether p* when and only when Transparency occurs; while epistemic accounts think that the question "Do you believe that p?" ask about the *subject's beliefs* both in cases in which Transparency occurs and in cases in which it doesn't. Secondly, the behavioural-expressivist account claims that the answer to the *transparent* question "Do you believe that p?" is an expressive episode of the pattern of an attitude that consists in a *judgment about whether p* (self-ascriptions of attitudes are characteristic of non-transparent third-person contexts); while epistemic accounts consider that the answer to the question "Do you believe that p?" is a first-person avowal that consists in a *self-ascription of attitude* regardless of whether Transparency occurs or not. Finally, the behavioural-expressivist account claims that the difference between first-person transparent contexts and third-person non-transparent contexts is *semantic* (i.e., whether the subject makes a judgement about a certain fact of the world or about her own mental states); while epistemic accounts consider that the difference between first-person transparent contexts and third-person non-transparent contexts is *epistemic* (i.e., whether the subject applies the special first-person procedure responsible for first-person epistemic self-knowledge or not).

1.5 Virtues of the behavioural-expressivist account of Transparency

The behavioural-expressivist account of Transparency is able to refute the arguments in favour of epistemic accounts because it explains the intuitions behind both Fernández's examples and Bar-on's semantic continuity argument without rendering a conclusion in favour of epistemic accounts. Firstly, it accounts for the lawyer's client and the priest examples because the questions "Do you believe that I am innocent?" and "Do you believe that God exists?" can be meant to be answered both in a *deliberative* and in a *self-ascriptive* way. When they are meant to be answered in a deliberative way, the intention of the questioner is to know whether one is innocent as a matter of fact (imagine a defendant asking her lawyer whether what she did constitutes a crime or not) or whether God exists as a matter of fact (imagine a priest asking a philosopher in search of reasons to rationally justify his belief in God). However, when the questions "Do you believe that I am innocent?" and "Do you believe that God exists?" are meant to be answered in a self-ascriptive way, the intention of the questioner is to know what the addressee *believes* about her own innocence or about the existence of God. In the examples described by Fernández, it is clear that the questions are meant to be answered in a self-ascriptive way because it is clear that the questioners are not interested in knowing whether one is innocent as a matter of fact or whether God exists as a matter of fact.

Therefore, in order to answer these questions (meant to be answered in a self-ascriptive way), the addressee can do different things depending on whether he appropriately understands the question and on whether he wants to follow through. On the one hand, if the addressee understands the question and if he wants to follow through, he will answer with a *self-ascription of attitude* issued by a third-person process of self-inspection (e.g., the third-person avowal "I believe that you are innocent" or the assertion "It is the case that I believe that God exists"). If the addressee hasn't made up his mind yet about what he believes because he didn't self-inspect himself about the issue before, he will make a self-inspective judgement on the basis of evidence about his own mental states (e.g., his passing thoughts, his feelings, his behaviour… about the innocence of the client or about the existence of God), thus forming a new second-order belief about his first-order attitude. By contrast, if the addressee already self-inspected himself about the issue in the past and he remembers the formed second-order belief, he will answer with the same self-ascription of attitude by which he concluded the self-inspective process in the past on the basis of no extra evidence about what he believes (e.g., "I believe that you are innocent" or "It is the case that I believe that God exists").

On the other hand, if the addressee doesn't understand the question or if he doesn't want to follow through, he will answer from the first-person deliberative perspective with *a judgement about the issue in question* (i.e., the innocence of the client or the existence of God). If the addressee hasn't made up his mind about the issue because he didn't deliberate yet, he will deliberate about the innocence of his client and about the existence of God on the basis of reasons and he will answer with a judgement about the issue in question (e.g., the assertion "You are innocent" or the first-person avowal "I believe that God exists"), thus giving rise to a new attitude (typically: belief, disbelief or suspension). If the addressee has already made up his mind because he has already deliberated about the issue and he still remembers, he will answer with the same judgement about the innocence of the client or about the existence of God by which he concluded his deliberation in the past made on the basis of no extra reasons about whether p (e.g., the assertion "You are innocent" or the first-person avowal "I believe that God does exist", etc.). Indeed, even if the questions are meant to be answered in a self-ascriptive way and the addressee mistakenly answers in a first-person deliberative way because he doesn't understand the question or because he doesn't want to follow through, any of these first-person deliberative answers will let the questioners know what the addressee *believes* about the innocence of the client or about the existence of God. For even if they are not self-ascriptions of attitude, they are expressive episodes (consisting in judgements about whether p) of the pattern of the first-order attitude formed by deliberation.

Secondly, the behavioural-expressivist account explains the intuitions behind Bar-on's semantic continuity argument because it explains both why my utterance "I believe that p" is true only when the ascription of belief "Jesús believes that p" made by a third subject about me is true and why I can use the utterance "I believe that p" as a premise in a sound argument while being its semantic content about my belief that p and not about the fact that p. On the one hand, the utterance "I believe that p" is true only when the utterance "Jesús believes that p" is true because of two reasons. Firstly, when "I believe that p" is issued as a *third-person avowal* of belief, it is a self-ascription of belief (i.e., it is an expressive episode of my second-order belief that I believe that p consisting in a judgement about the fact that I *believe* that p), and so, its truth-conditions are the same as the *truth-conditions* of the ascription of belief "Jesús believe that p" made by a third-person subject about me. Secondly, when "I believe that p" is issued as a *first-person avowal*, it is an expressive episode of my belief that p (consisting in the judgement that p is the case), and so, it is the *truth-maker* of the ascription of belief "Jesús believes that p" made by a third-person about me. On the other hand, if I can use "I believe that p" (with the semantic content of my belief that p and not with the semantic content of the fact that p) as a premise in a sound argument whose premises transmit their truth to the conclusion is because "I believe that p" can be issued as a third-person avowal (i.e., as a self-ascription of belief) whose semantic content is my belief that p. Therefore, the behavioural-expressivist account can explain the phenomena pointed out by Bar-on because avowals have different meanings (i.e., they can be about me or about the world) depending on whether they are issued from the first-person deliberative perspective or from the third-person self-inspective perspective.

Moreover, the behavioural-expressivist account of Transparency not only refutes the arguments in favour of epistemic accounts, it also explains the phenomenon of Transparency better than epistemic accounts because it is able to explain in a better way 1) how the deliberation about whether p by which subjects answer the transparent question "Do you believe that p?" works 2) without facing the objection that Transparency doesn't explain first-person epistemic self-knowledge of our already held beliefs (as epistemic accounts face). Firstly, it has been argued (García, 2019a) that the behavioural-expressivist account of Transparency provides a more plausible explanation of the deliberation about whether p behind Transparency than epistemic accounts because it claims that the episode of expression that concludes the deliberation about whether p is a *judgement about whether p* that originates a new attitude (typically: belief, disbelief or suspension) instead of a *self-ascription of attitude* (as epistemic accounts of Transparency claim). Here is a plausible description of how that deliberation goes:

> "Take an agent who, deliberating as to whether p is the case, utters the following open-ended string of sentences: "p is the case …. No, not p, q is the case …. Wait a minute, r is the case …." This is a clear instance of ongoing deliberation as to whether p, in which the agent's mind is not yet settled as to what to believe. Consider now an agent (perhaps the same agent at a different time) who, in a similar deliberative context as to whether p, utters the following open-ended string of sentences: "I believe that p …. No, not p, I believe that q …. Wait a minute, I believe that r …." How should we describe what is going on here? Is this also an instance of ongoing deliberation as to whether p; or is it rather a case in which the agent has not settled on a judgement about the belief he or she holds? What the latter option means is that the agent's mind is settled as to whether p, and it is only the judging of what his or her mind is like that is still unsettled. However, in the deliberative context under consideration, it is more natural to view the open-ended string of sentences "I believe that p …. No, not p, I believe that q …. Wait a minute, I believe that r …." as what having an unsettled mind as to whether p consists in, by analogy with the open-ended string of sentences: "p is the case …. No, not p, q is the case …. Wait a minute, r is the case ….". (García, 2019a, p. 141)

In this description, first-person avowals (e.g., "I believe that p") and assertions about whether p (e.g., "p is the case") seem to play the same linguistic role insofar as both are exchangeable episodes of the subject's ongoing deliberation about whether p. So, if they play the same role insofar as they all are episodes of an ongoing deliberation about whether p, it is plausible to think that they will still play that role when they are issued as the *conclusion* of the ongoing deliberation about whether p. Therefore, to claim (as the behavioural-expressivist account of Transparency does) that subjects conclude their deliberations about whether p with a judgement about whether p that is an expressive episode of an attitude and that can take the linguistic form of a first-person avowal (e.g., "I believe that p") or the linguistic form of an assertion (e.g., "p is the case") is more plausible than to claim that subjects conclude their deliberations about whether p with a self-ascription of attitude (as epistemic accounts of Transparency do). For, under this hypothesis, there would be a change of subject matter between the ongoing deliberation (which is about whether p) and the reached conclusion (which would be about whether I believe that p). As a result, the behavioural-expressivist account explains in a more plausible way the deliberation about whether p behind Transparency.

Secondly, the objection that Transparency is not able to explain cases of first-person epistemic self-knowledge of our already held beliefs doesn't apply to the behavioural-expressivist account. On the one hand, the behavioural-expressivist account thinks that Transparency is a semantic phenomenon that has nothing to do with first-person epistemic self-knowledge (*knowing that*). On the other hand, as it is going to be argued in the next chapter, the behavioural-expressivist account considers that first-person epistemic self-knowledge is not a genuine phenomenon, and so, that trying to explain first-person self-knowledge in an epistemic way (i.e., as true warranted belief) is the first error made by epistemic accounts of Transparency. By contrast, the behavioural-expressivist account considers that first-person self-knowledge is an expressive phenomenon that has to do with the way in which one exercises the ability to express one's own mental states (*knowing how*).

Therefore, the state of the art on Transparency shows that there are already reasons to prefer the behavioural-expressivist account of Transparency over epistemic accounts. For the behavioural-expressivist account manages to explain Transparency in a more plausible way (i.e., considering that the deliberation involved in Transparency is about the same subject matter as the judgement that concludes that deliberation: about whether p) while avoiding the objection adduced against epistemic accounts (i.e., that they cannot explain first-person epistemic self-knowledge of already held attitudes) and while refuting the arguments in favour of epistemic accounts of Transparency (i.e., Fernández's examples and the semantic continuity argument). In the next chapter, behavioural expressivism is going to be explicated in detail. To do that, behavioural expressivism is going to be compared with neo-expressivism in regard to two issues: their takes on expression and their takes on self-knowledge.

2. Neo-Expressivism vs. Behavioural Expressivism

The concept of expression has been used by neo-expressivists (Bar-on, 2004, 2013; Finkelstein, 2003) to account for a variety of mental phenomena, such as consciousness, first-person self-knowledge, or the asymmetry between first-person and third-person avowals. Neo-expressivists usually differentiate their accounts from a group of accounts known as “classical expressivism”. According to classical expressivism, there are certain kinds of sentences that don’t have truth-value or propositional content because they are expressions of mental states *rather than* assertions of facts of the world. For instance, classical expressivism about *moral judgements* (e.g., Ayer, 2001; Stevenson, 1944) considers that moral statements (e.g., “Stealing is bad”) are expressions of evaluative attitudes (i.e., approval or disapproval) rather than assertions of moral facts with a particular truth-value and propositional content. Or classical expressivism about *first-person avowals* (e.g., Wittgenstein in *Philosophical Investigations*, under some readings —e.g., P.F. Strawson, 1954; Malcolm, 1954—) considers that first-person avowals (e.g., “I have a pain in my leg”) are expressions of mental states rather than assertions of psychological facts with a particular truth-value and propositional content. Thus, classical expressivism is understood as the thesis that there are certain kinds of sentences (e.g., moral judgements, first-person avowals…) that are linguistic expressions of mental states that don’t have any truth-value or propositional content because they don’t assert anything at all (like a gesture of disgust or a cry of pain).

Neo-expressivism is developed as an alternative to classical expressivism. Neo-expressivist accounts consider that being an expression is not incompatible with having truth-value or propositional content, and so, that it shouldn’t be ruled out that quickly that first-person avowals are expressions of mental states *and* assertions of mental states at once. True, there are expressions of mental states that don't assert anything, and so, that don't have any truth-value or propositional content (e.g., a gesture of disgust or a cry of pain). But there are other expressions of mental states that are expressions and assertions at once due to the fact that they have a linguistic form, and so, they have a particular truth-value and propositional content. First-person avowals (e.g., "I believe that p") and other linguistic expressions (e.g., "p is the case") are examples of expressions with truth-value and propositional content.

There is an important similarity between neo-expressivist accounts and the behavioural-expressivist account regarding the nature of first-person avowals. Both neo-expressivism and behavioural expressivism claim, against classical expressivism, that first-person avowals (e.g., "I believe that p", "I have a pain in my leg", etc.) are *truth-evaluable linguistic expressions*. However, there are also two fundamental differences between them. Firstly, neo-expressivist accounts endorse a *relational view of expression* according to which mental states and their characteristic set of expressions are two different items related in some way. By contrast, the behavioural-expressivist account endorses a *non-relational view of expression* according to which expressions are episodes of mental states rather than further items related in some way (more about this below). Secondly, neo-expressivist accounts understand first-person avowals as *self-ascriptions* of mental states that express the self-ascribed mental state itself (as opposed, for instance, to express only a second-order belief). The truth-evaluable character of those self-ascriptions *plus* the fact that they are expressions of the self-ascribed mental state itself are supposed to set the basics to explain *first-person epistemic self-knowledge*. By contrast, the behavioural-expressivist account considers that first-person avowals are *not self-ascriptions* of mental states (even though they are truth-evaluable expressions of mental states) so that *first-person self-knowledge*, understood as an epistemic phenomenon (i.e., true warranted belief), doesn't exist; first-person self-knowledge is always *expressive* self-knowledge (more about this below). These two key differences between neo-expressivist accounts and the behavioural-expressivist account are going to be explicated in the following sections (i.e., in sections 2.1 and 2.2, respectively).

2.1 Relational views vs. non-relational views of expression

According to relational views of expression, expressions are the subset of *manifestations* of a subject (e.g., tears, smiles, facial movements, demeanour, interjections, utterances…) that acquire a particular *expressive content* because of the relation that they have with a different item: a *mental state* (e.g., happiness, sadness, nervousness, pain, belief, intention, desire, hope, etc.); as opposed to the subset of manifestations of a subject that don't acquire any expressive content because they are not related to any mental state (e.g., tears of allergy). Thus, in this view, S's *expressive manifestations* and S's *mental states* are considered to be two different sets of items related in some way. Insofar as expressions and mental states are considered to be different sets of items, they are supposed to be *detachable* from each other without ceasing to be the *kind of things* that they are. So, a mental state can allegedly take place in a subject without its characteristic set of expressions, and *vice versa*, the set of expressions characteristic of a mental state can allegedly take place in a subject without the mental state taking place in the subject as well. By contrast, non-relational views of expression consider that expressions are the subset of manifestations of a subject that have *intrinsic expressive content* (a property that depends on the kind of thing that a manifestation is and not on a relation to a further item); as opposed to the subset of manifestations of a subject that don't have any intrinsic expressive content (e.g., tears of allergy). Therefore, non-relational views of expression consider that the set of expressions characteristic of a mental state (e.g., tears, certain facial expressions, certain bodily postures, a certain demeanour, saying "I'm sad", etc.) and the corresponding mental state itself (e.g., sadness) are *one and the same item*, meaning that mental states are nothing over and above a particular set of expressions. Insofar as the set of expressions characteristic of a mental state and the corresponding mental state itself are considered to be one single item, they are not considered to be detachable from each other while still being the same *kind of thing* that they currently are. Thus, a mental state and its corresponding set of expressions are not supposed to be able to take place separately in a subject.

In this section, it is going to be argued that the non-relational view of expression endorsed by the behavioural-expressivist account is preferable over the relational views of expression, including the relational view of expression characteristic of neo-expressivist accounts. Firstly, it is going to be argued that relational and non-relational views of expressions are dependent on the way in which it is understood that expressions are evidence of mental states and the relational view characteristic of neo-expressivist accounts is going to be explained. Secondly, the non-relational view of expression endorsed by the behavioural-expressivist account is going to be explicated in detail. And thirdly, it is going to be argued that the non-relational view of expression characteristic of behavioural expressivism explains cases of *pretence* and *dissimulation* better than relational views of expression.

2.1.1 Expression as evidence of mental states

Expressions are evidence of mental states. Depending on the way in which it is understood that expressions are evidence of mental states, it follows either a non-relational view of expression or a relational view of expression. There are two types of evidence: symptomatic evidence and criterial evidence. In turn, it is discussed whether criterial evidence is defeasible or indefeasible. In this section, it is going to be shown that relational views of expression assume either that expressions are *symptoms* of mental states (i.e., relational views of a causal type) or that expressions are *defeasible criteria* of mental states (i.e., relational views of a mereological or constitutive type, characteristic of neo-expressivism) and that from endorsing the idea that expressions are *indefeasible criteria* of mental states follows a non-relational view of expression. These relations will be important to understand the precise differences between the neo-expressivist and the behavioural-expressivist conceptions of mental states.

There are two types of evidence: *symptoms* and *criteria*[10]. X is evidence of Y of the symptomatic type when X and Y are two different facts or events and when there is a stable *empirical correlation* between the occurrence of X and the occurrence of Y. Since the only connection between X and Y is the occurrence of a stable empirical correlation between them (e.g., having an external causal connection[11], being effects of a common cause, etc.), X won't

[10] This distinction was explicitly made and discussed by Wittgenstein in *The Blue and Brown Books* (1958, pp. 24-25).

[11] One can distinguish between external and internal causal connections. External causal connections take place between two *different* and *independent* facts or events, whereas internal causal connections take place within two different aspects of the *same* fact or event or within an aspect of a fact or event and the whole fact or event.

be (symptomatic) evidence of Y in those particular cases in which, due to abnormal conditions, the empirical correlation is broken because X takes place without Y taking place as well. However, as far as the correlation continues to occur at a *general level*, X will continue to be *prima facie* symptomatic evidence of Y, and so, X will continue entitling us to predict the *likeliness* of the occurrence of Y. For instance, a barometer indicating low pressure is evidence of rain of the symptom type because barometers indicating low pressure and occurrences of rain are two different events and because there is a stable empirical correlation between them. Since the connection between the barometer indicating low pressure and the rain is based on a stable empirical correlation, a barometer indicating low pressure won't be symptomatic evidence of rain in those particular cases in which, due to abnormal circumstances (e.g., the barometer is broken), the empirical correlation between the barometer and the rain breaks down. However, as far as the empirical correlation between barometers indicating low pressure and occurrences of rain continues taking place at *a general level*, barometers indicating low pressure will continue being *prima facie* symptoms of rain and they will continue entitling us to predict the likeliness of the occurrence of rain.

By contrast, the notion of criteria refers to a stronger type of evidence based on a *conceptual connection*. X is evidence of Y of the criterial type when there is a conceptual connection between X and Y so that X and Y are connected by the kind of things that they are: a *fully-fledged fact or event* (Y) and a particular *aspect* (X) of that very same event. So, while symptoms are facts or events that inform us of the likeliness of the occurrence of *different* facts or events, criteria are *aspects* of fully-fledged facts or events which can give us *perceptual access* to the facts or events that they are criteria of. For instance, the small number of raindrops (relatively) that I see and hear falling on the street through my window when it is raining is criterial evidence of the fact that it is raining because those few raindrops are *aspects* of the event of rain as a whole, aspects that can enable my *perceptual access* to the event of rain itself. Furthermore, insofar as those few raindrops are aspects of the fully-fledged event of rain (which includes multiple additional aspects; e.g., an enormously bigger number of raindrops, a cloud of water vapour condensing itself into liquid, etc.), the event of rain and the few raindrops that I perceive through my window are *conceptually connected* by the kind of things that they are (i.e., a fully-fledged event of rain and a particular aspect of that event —the raindrops that I see—).

The notion of criterial evidence might be understood in two different ways depending on whether it is understood as defeasible evidence or as indefeasible evidence. On the one hand, those who understand criteria as *defeasible evidence* (e.g., Albritton, 1959; Gaynesford, 2002; Lycan, 1971; Shoemaker, 1963; Witherspoon, 2011; Wright, 1984) consider that X is a criterion of Y if and only if X is *always and necessarily* conceptually connected to Y so that X is *always and necessarily* (criterial) evidence of Y.[12] Of course, these authors are aware of the fact that sometimes, as a matter of fact, X takes place without Y taking place as well so that X doesn't present an aspect of Y nor can it enable perceptual access to Y (for, again, Y doesn't take place at all). However, instead of explaining these cases as cases in which X is *not* a criterion of Y because X and Y take place separately, they claim that these are cases in which X is still a criterion of Y (for X is still conceptually connected to Y) with the qualification that X is a *defeated* criterion of Y (for Y doesn't take place as a matter of fact in this particular case). Hence the distinction between a defeated criterion (i.e., X takes place without Y taking place) and an undefeated criterion (i.e., both X and Y take place). Let me explain this with an example. Imagine that, in order to deceive me, my upstairs neighbour recreates an aspect of the event of rain (e.g., a few drops of water falling on the street) by meticulously throwing water from his window with a shower so that I have a perceptual experience indiscernible from the perceptual experience that I would have if it were raining and I were looking outside from my window. If you consider the criterion of rain to be defeasible, you consider that the criterion of rain is present here. For the *same aspect* of the event of rain that I perceive from my window when it is raining (i.e., some drops of water falling on the street) is being perfectly recreated by my neighbour in order to deceive me. That's why, in this case, I can be misled into thinking that it is raining by looking through the window. True, the criterion is *defeated* in this case because the event of rain doesn't take place as a matter of fact, but the drops of water falling on the street are still *criterial evidence* of rain because they are *conceptually connected* to events of rain: every event of rain contains (among many other things) a few drops of water falling on the street in the way that I see from my window, and in this sense, those few drops of water coming from my neighbour's shower are aspects of the world of the *same kind* as

[12] Shoemaker formulates this idea quite clearly:

> "A test of whether something is one of the criteria for the truth of judgements of a certain kind is whether it is conceivable that we might discover empirically that it is not, or has ceased to be, evidence in favour of the truth of such judgements. If it is evidence, and it is not conceivable that it could be discovered not to be (or no longer be) evidence, then it is one of the criteria. If so and so's being the case is a criterion for the truth of a judgment of φ-identity, the assertion that it is evidence in favour of the truth of the judgment is necessarily (logically) rather than contingently (empirically) true. We know that it is evidence, not by having observed correlations and discovered empirical generalizations, but by understanding the concept of a φ and the meaning of statements about the identity of φ's." (Shoemaker, 1963, pp. 3-4).

some of the aspects that compose any fully-fledged event of rain (particularly, of the same kind as a few *raindrops* falling on the street).

I think that to allow for the possibility of a criterion being *defeated*, the conceptual connection characteristic of criteria has to be understood as *forming part* of an internal relation between X and Y which has a *dual character*: it is conceptual and empirical at once[13]. For this idea could help to explain the possibility of the internal relation being *somehow broken* (when X is defeated and X doesn't present an aspect of Y because Y doesn't take place) without X and Y being *totally detached* (for, as criterial evidence, X still *presents an aspect of the world* that is conceptually connected to Y). On the one hand, insofar as the internal relation between X and Y is *conceptual*, X is conceptually related at the same time with all the fully-fledged facts or events that X *can be* an aspect of (e.g., Y, Y', Y'', etc.); so X always and necessarily is *criterial evidence* of those facts or events (e.g., Y, Y', Y'', etc.). For instance, insofar as a few drops of water falling on the street (X) can be an aspect both of fully-fledged events of rain (Y) and of fully-fledged events of someone throwing water with a shower to feign rain (Y'), a few drops of water falling on the street are considered to be an aspect of the world that is *always and necessarily* conceptually connected both to events of rain (Y) and to events of someone throwing water with a shower to feign rain (Y'). For X (i.e., a few drops of water) is connected to Y (i.e., rain) and to Y' (i.e., someone feigning rain) by the kind of things that X, Y and Y' are: an aspect of fully-fledged events (X) and two of the fully-fledged events (Y and Y') of which X can be an aspect. As a result, X is *conceptually undetachable* from the facts or events of which it can be an aspect (Y, Y', Y'', …). On the other hand, insofar as the internal relation between X and Y is *empirical* as well, the particular case needs to be taken into account in order to find out whether X is a *defeated* or an *undefeated* criterion in regard to Y. For instance, if it happens in a particular case that the drops of water falling on the street (X) are an aspect of the event of rain because it is raining (Y), then X is an undefeated criterion in regard to Y (i.e., the rain) and a defeated criterion in regard to Y' (i.e., someone feigning rain). By contrast, if it happens in a particular case that the drops of water falling on the street (X) are an aspect of the event of someone feigning rain (Y'), then X is an undefeated criterion in regard to Y' (i.e., someone feigning rain) and a defeated criterion in regard to Y (i.e., the rain). As a result,

[13] Most accounts of defeasible criteria focus only on the epistemic and the semantic elements of criterial evidence, neglecting so the ontological aspect of it (which is just the other side of the coin). Since I will trace some lines between the way to understand expressive evidence of mental states and the way to understand the nature of mental states, I will try to reconstruct here the ontological view of criteriological relations that I think the notion of defeasible criteria entails.

the criterion X and the facts or events of which X can be an aspect (Y, Y', Y'', ...) are *empirically detachable*.

Two interesting ideas seem to follow from the notion of defeasible criteria. Firstly, from the idea that the criterion X and the facts or events Y, Y' or Y'' are capable of being empirically detached (defeated criterion) without ceasing to be conceptually connected seems to follow the idea that the *kind of thing* that X is doesn't depend on whether (as a matter of fact) it is an aspect of the fully-fledged fact or event Y, Y', or Y''. For instance, if a few drops of water falling on the street are considered to be a criterion (either defeated or not) of the event of rain, of the event of someone feigning rain, and of the event of a fire-fighting plane dropping water from the sky (all at the same time), those few drops of water falling on the street will always be the same kind of thing (the criterion X), regardless of whether, in a particular case, they are an aspect of the event of rain, an aspect of the event of someone feigning rain, or an aspect of the event of a firefighting-plane dropping water from the sky. In all these cases, those few drops of water falling on the street (the criterion X) are an *aspect of the world* that can conceptually take part in the variety of facts or events of which it is criterion (Y, Y', Y''). That's why, for instance, we are supposed to be able to find criterial evidence of rain in a variety of facts or events that are different from rain but which contain an aspect of the world that is conceptually connected to the event of rain; i.e., a few drops of water falling on the street can be found in the event of someone feigning rain, in the event of a fire-fighting plane dropping water from the sky, in the event of a clumsy person watering the plants on her balcony, etc. Therefore, the criterion X and the fully-fledged facts or events Y, Y' or Y'' of which X can be an aspect (i.e., criterion) are conceived here as *two different sets of items* insofar as the criterion X is supposed to be detachable from the fully-fledged facts or events Y, Y' or Y'' without changing the kind of thing that it is: an *aspect of the world* that can conceptually take place in the variety of facts or events of which it is a criterion (i.e., the criterion X).

Secondly, from the idea that the criterion X and the fact or event Y are different items seems to follow the idea that X (when it is an undefeated criterion of Y) can provide *indirect* perceptual access to Y rather than *direct* perceptual access. Indeed, if the kind of thing that the criterion X is (e.g., a few drops of water falling on the street) doesn't depend on the particular fact or event of which X is an aspect on a given occasion (e.g., the event of rain, the event of someone feigning rain, the event of a fire-fighting plane dropping water…), it follows that *the content of my perceptual experience* is always the same when I perceive the criterion X, regardless of the particular fact or event of which X is an aspect in the particular case. As a result, the criterion X can provide only *indirect* perceptual access because it acts as a *mediating entity* between the perceiving subject and the fact or event of which it is an aspect in the particular case. Indeed, it acts as a mediating entity because the criterion X is always the *same kind of thing*, regardless of the fact or event of which it is an aspect on a given occasion, and so, the content of my experience when I perceive the criterion X is always the same regardless of the particular fact or event of which it is an aspect. For instance, it follows from the notion of defeasible criteria than when I have perceptual access to the event of rain (Y) by looking at the drops of water falling on the street (X), the content of my experience must be the same as when I have perceptual access to the event of someone feigning rain (Y') by looking at the drops of water falling on the street (X). In both cases, the content of my experience must be the same because I perceive the same kind of thing: a few drops of water falling on the street (the criterion X). And from the fact that the content of my experience is the same both when I perceive some water drops (X) of rain (Y) and when I perceive some water drops (X) of someone feigning rain (Y') follows that the criterion X is a mediating entity between me and the perceived fact or event (i.e., the rain or the feigned rain, depending on the case).

On the other hand, those who understand criteria as *indefeasible* evidence (McDowell, 1998) consider that the fact that X is criterial evidence of Y *in certain cases* doesn't involve that X and Y are always and necessarily conceptually connected, and so, it doesn't involve that X is always and necessarily (criterial) evidence of Y. When X is criterial evidence of Y, X and Y are conceptually connected because X is an actual *aspect* of the event or fact Y, and so, X can enable *perceptual access* to Y. Since X is an actual aspect of Y in this case, X is an *indefeasible* criterion of Y: if X takes place as criterion of Y, Y necessarily takes place as well. By contrast, when X and Y are not conceptually connected because X is not an aspect of the event or fact Y, X is not criterial evidence of Y nor can it enable perceptual access to Y. Since X is not an aspect of Y in this case, X can take place without Y taking place as well. However, that doesn't mean that X is a *defeated* criterion of Y but that X is *not a criterion* of Y at all in this case precisely because X is not an aspect of Y in this case. Thus, criterial evidence is considered to be indefeasible here because it is impossible that X takes place as criterion of Y without Y taking place as well. Wherever X is criterial evidence of Y, X is conceptually connected to Y because it is an actual aspect of the occurrent fact or event Y, and so, it always can enable perceptual access to Y. Let's continue with the example of my fastidious neighbour. If it seems to me that it is raining outside when what I'm actually seeing are drops of water falling from my neighbour's shower, then I'm wrong and I may discover my mistake later on (e.g., taking a closer look). But the *criterion of rain* doesn't take place here at all because I'm not perceiving an actual *aspect of the event of rain* insofar as it is not raining. Quite differently, unbeknown to me, I see the *criterion of my neighbour's attempt to feign rain* (since an actual aspect of that event is what I see through the window) even though I mistakenly take it for the criterion of rain. So, a few drops of water falling on the street can either be an indefeasible criterion of rain (when they are an actual aspect of the event of rain) or an indefeasible criterion of my neighbour's attempt to feign rain (when they are an actual aspect of that event), but they cannot be both at the same time. In this case, the drops of water that I see are an indefeasible criterion of my neighbour's attempt to feign rain and not an indefeasible criterion of rain (even if I can't tell the difference from my current epistemic situation) because they are an aspect of the former event and not an aspect of the latter. Then, from my *current epistemic limitation* to tell the difference between those two criteria (i.e., water drops of actual rain and water drops of fake rain) doesn't follow the *ontological identity* of the two criteria themselves (i.e., water drops of actual rain and water drops fake rain). By contrast, the two criteria are very different ontologically because they are aspects of very different events (i.e., rain and fake rain), as I might discover later on by changing my current epistemic position for another more advantageous one (e.g., by standing up from the sofa and taking a closer look through the window).

There are two interesting ideas that follow from the notion of indefeasible criteria. Firstly, since the possibility of a defeated criterion is ruled out here, the conceptual connection between the criterion X and the fact or event Y cannot be understood as forming part of an internal relation between X and Y with *both* an empirical and a conceptual character. By contrast, the criterion X and the fact or event Y must be considered as *one and the same item* because their conceptual connection is nothing over and above their being two different *perspectives* from which one and the same item can be considered: the criterion X is an aspect of Y and Y is the fully-fledged fact or event of which X is an aspect. As a result, that conceptual connection between X and Y cannot occur when X takes place separately from Y. For when X takes place separately from Y, X is not an actual aspect of Y, and so, X is not a criterion of Y in these cases. For instance, the event of rain can be considered from the perspective of a fully-fledged phenomenon that includes as many aspects as one can locate in the event (e.g., a cloud condensing into water, an enormous amount of drops of water, a rainbow, etc.); or it can be considered from the perspective of one of those particular aspects (e.g., some drops of water seen from my window). Each one of these aspects is a criterion of rain. The proof of the fact that the criterion X and the fact or event Y are not different items, but only two different perspectives from which a single item can be considered, is that the criterion X cannot be detached from the event Y without changing the *kind of thing* that it is. For instance, if we think of the water drops of rain falling on the street as separated from the event of rain (e.g., as drops of water coming from my neighbour's shower), we think of the drops of water as not being *raindrops* (i.e., as not being criterion of rain insofar as they are not actual aspects of the event of rain) but drops of water of some different event (e.g., as criterion of my neighbour's attempt to feign rain). The idea here is not that the water drops of rain cannot be separated from the event of rain *as water drops*, but that they cannot be separated from the event of rain as *raindrops* (i.e., as criterion of rain). For as soon as they are considered as a separated item from the event of rain, they stop being *raindrops* because they stop being an actual aspect of the event of rain (i.e., they stop being criterion of rain). Thus, the raindrops and the event of rain are conceptually connected only insofar as they are two perspectives from which one and the same item can be considered: the fully-fledged event of rain and one of its aspects (e.g., some raindrops).

Secondly, from the idea that the criterion X and the fact or event Y are one and the same item follows the idea that the criterion X can enable *direct* perceptual access to Y rather than *indirect* perceptual access. For it follows that the content of my experience is *different* when I perceive the criterion X (i.e., some drops of water) of Y (i.e., the event of rain) and when I perceive the criterion X' (i.e., some drops of water) of Y' (i.e., the event of my neighbour feigning rain), even if they are similar perceptions in appearance. Imagine that I perceive from my window some drops of water falling on the street. Even if I may not be able to distinguish from my current epistemic position whether I perceive the criteria of rain or the criteria of my neighbour's attempt to feign rain, the content of my experience is *different* in both cases because in one case I perceive the criteria of rain (i.e., actual raindrops) and in the other case I perceive the criteria of my neighbour's attempt to feign rain (i.e., water drops of fake rain). As a result, it cannot be said that the criterion of rain is a mediating entity between the perceiving subject and the perceived fact or event; instead, the criterion can provide *direct* perceptual access to the perceived fact or event that it is an aspect of. In one case, the content of my perceptual experience is the criterion of rain (i.e., a few actual raindrops), and so, I can directly perceive the event of rain in that criterion of rain; in the other case, the content of my perceptual experience is the criterion of someone feigning rain (i.e., a few water drops of fake rain), and so, I can directly perceive the event of fake rain in that criterion of fake rain. True, it might be that I am not able to distinguish at this particular moment whether I perceive raindrops or water drops coming from my neighbour's shower, but that doesn't mean that the content of my experience must be the same (i.e., ontologically identical) in both cases. What that means is that I have to change my current epistemic position (e.g., standing up from the sofa to take a closer look at what's happening outside) to be able to find out which is the actual content of my current perceptual experience; that is, to be able to find out which event I perceive through my window in this case (rain or my neighbour's attempt to feign rain). So, to doubt which event I perceive in this particular case is tantamount to doubt which is the true content of my current perceptual experience (i.e., raindrops or drops of water from my neighbour's shower); and to find out which event I perceive in this particular case is tantamount to find out which is the true content of my current perceptual experience (i.e., raindrops or drops of water from my neighbour's shower).

There are epistemic reasons to think that to understand the notion of criteria as indefeasible evidence is preferable to understanding the notion of criteria as defeasible evidence. When we say that X is evidence of Y, what we want to say is that X counts in favour of the *empirical occurrence* of Y, either by showing that Y is more likely (if X is a symptom of Y) or by presenting an aspect of Y itself (if X is a criterion of Y). For only if X counts in favour of the *empirical occurrence* of Y, X can warrant the belief that Y is likely or the belief that Y is the case, which is exactly what evidence is supposed to do in order to provide epistemic knowledge (i.e., true warranted belief). Then, if evidence must count in favour of the empirical occurrence of something in order to act as evidence, and if criteria is supposed to be a type of evidence, it is difficult to see which are the advantages of understanding the conceptual connection characteristic of criteria in a way that allows the possibility of X taking place as a criterion of Y without Y empirically taking place as well (defeasible notion of criteria). Especially, when an alternative way to understand the conceptual connection characteristic of criteria is available: X is criterial evidence of Y only when X is an actual aspect of Y, and so, it is impossible that X can take place *as criterial evidence* of Y without Y empirically taking place as well (indefeasible notion of criteria).

As it was said before, different views of expression follow depending on how it is understood that expressions are evidence of mental states. Particularly, from considering that expressions are symptomatic evidence of mental states, it follows a cause-and-effect relational view of expression; from considering that expressions are defeasible criteria of mental states, it follows a mereological or constitutive relational view of expression; and from considering that expressions are indefeasible criteria of mental states, it follows a non-relational view of expression. Firstly, from the idea that expressions are *symptomatic evidence* of mental states follows the idea that between expressions and mental states there is a *cause-and-effect* type of relation (e.g., Armstrong, 1995, pp. 175-190; Putnam, 1975, Vol. 2, pp. 325-341). According to this cause-and-effect relational conception, some manifestations of S (e.g., tears, smiles, demeanour, interjections, utterances, actions, etc.) acquire a particular expressive content (e.g., sadness, happiness, excitement, belief, hope, etc) when they are the *causal effect* of a different and independent kind of item: a mental state of S. Then, the expressive content (if any) of S's manifestations is considered to be the result of an external causal relation between two sets of different and independent items: S's manifestations and S's mental states. Since these cause-and-effect relations are considered to be external, the *same kind* of manifestation (e.g., some tears) is supposed to be able to acquire different expressive contents, or no expressive content at all, depending on its causal relations. Thus, when the manifestation is not the effect of a mental state of S but of something else, it doesn't acquire any expressive content (e.g., tears caused by an allergy). And when the manifestation is the effect of a mental state of S, it does acquire the expressive content of that particular mental state (e.g., tears caused by sadness express sadness and tears caused by happiness express happiness). Moreover, when a manifestation of S has expressive content (i.e., when it is caused by a mental state), it is considered to enable *indirect non-perceptual access* (e.g., inferential access) to the mental state that it expresses.

Secondly, from the idea that expressions are defeasible criteria of mental states follows the idea that between expressions and mental states there is a *mereological* or *constitutive* kind of relation. According to mereological or constitutive relational views, some manifestations of S have a certain expressive content because they are aspects or components of a whole mental state, and so, they are *internally related* to that whole mental state (i.e., a totality formed by a set of aspects or components) in the same way as a part is supposed to be related to the whole (if one endorses a defeasible notion of criteria): with an internal relation that has both a conceptual and an empirical character. On the one hand, S's expressive manifestations (e.g., "I have a headache") are *conceptually connected* to a particular mental state or totality (e.g., headache) because they are always and necessarily the kind of things that can be aspects or components of that particular mental state or totality, and so, they always and necessarily express the mental state or totality of which they can be an aspect or component (e.g., "I have a headache" always expresses headache). Then, S's expressions are *conceptually undetachable* from their corresponding mental states. On the other hand, S's expressive manifestations can be *empirically related* to a particular totality or mental state of S because they are aspects or components that can (as a matter of fact) take place within their corresponding totality or mental state (e.g., to say "I have a headache" when I actually have a headache) or take place separately from their corresponding totality or mental state (e.g., to say "I have a headache" pretending to have a headache). Then, S's expressions and S's mental states are *empirically detachable* because (as a matter of fact) both can take place separately. As a result, S's expressions and S's mental states are considered to be two different sets of items related in the same way that a part is supposed to be related to the whole (if one endorses a defeasible notion of criteria): even though expressions and mental states are always conceptually connected (i.e., expressions are always a possible part of a particular mental state as a whole), they can take place separately (i.e., expressions can be detached from their mental state). As a result, S's expressions are supposed to provide *indirect perceptual access* to S's mental states (which are considered to be additional items); i.e., perceptual access to S's mental states *through* S's expressions.

Neo-expressivism endorses a relational view of expression that seems to be of this *mereological or constitutive* kind, at least in Bar-on's version of neo-expressivism. According to Bar-on, mental states are *conditions* of a subject (Bar-on, 2004, p. 424) which cannot be reduced to a repertoire of expressions (Bar-on, 2004, p. 421). However, even if mental states cannot be reduced to a repertoire of expressions, expressions are *components* of the mental states that they express. So, expressions are aspects or components of a totality or condition of the subject; i.e., of a mental state as a whole. To explain her view, Bar-on suggests the analogy of a maple tree and its branches:

> "I can see the maple tree in my yard by seeing a characteristic component of it—say, one of its branches. The branch could be severed and separated from the tree, so it is possible for me to see the branch without seeing the tree. But that doesn't change the fact that *if* the tree is there, still attached to the branch, I can see the tree by seeing the branch. In other words, *if* I were to see the tree—which requires that the tree be there so as to be seen—it would be by seeing the branch that I would see it. As suggested by Green, we could perhaps think of natural expressions as exhibiting characteristic

> components of the states they express, so we can sometimes see the relevant state by seeing a characteristic component of it." (Bar-on, 2004, p. 228).

Later, Bar-on expands this idea to non-natural expressions, that is, to expressions that are culturally learned during the socialization process of the individual (such as, linguistic expressions):

> "On the expressivist story I offered there, mental terms are handed down to learners as components of new forms of expressive behavior. If so, we can regard them as keyed to the *conditions that users of the terms perceive in subjects' expressive behavior*. What we take subjects to express, however, are conditions the subjects are in, not states that are in the subjects. As expressing is something we take subjects to be doing, mental conditions are taken to be conditions of subjects, or conditions they are in, rather than states *inside* them." (Bar-on, 2004, p. 424).

Following the analogy of the maple tree, S's expressions are normally *components* of S's mental states just as maple-tree branches are normally components of a maple tree. As a result, S's expressions *normally* enable us to have indirect perceptual access to S's mental states; e.g., we can normally perceive someone's sadness through her tears or through her facial expressions[14]. So, S's mental states can be *normally* perceived through S's expressions just as a maple tree can be *normally* perceived through one of its branches. However, we can't always perceive S's mental states in S's expressions. Just as we can perceive a maple tree branch without perceiving a maple tree when the branch is severed from the trunk, Bar-on argues that sometimes we perceive S's expressions without perceiving S's corresponding mental state

[14] That Bar-on endorses an *indirect* view of our perceptual access to mental states is not only coherent with the way she understands the relation between expressions and mental states, it is also implied in the idea that the perceptual content of seeing a tree in one of its branches (mutatis mutandis, of seeing a mental state in one of its expressions) might be insufficient to give us perceptual access to the tree itself (mutatis mutandis, to give us perceptual access to S's mental state itself):

> "If we think of the presence of the behavior as what enables us to perceive someone's being in a mental state, this suggests that there is something more, or something else, to her being in the state than engaging in the perception-enabling behavior. (Analogy: if we see the tree by seeing one of its branches, seeing the tree is not the same thing as seeing the branch.)." (Bar-on, 2004, p. 423).

because S doesn't have such a mental state[15]. In this type of cases, S's expressions are still expressions of a mental state in a conceptual way (just as a severed maple tree branch is still a branch of maple tree in a conceptual way), but S's expressions don't express a mental state *of* S because, as a matter of fact, S doesn't have that mental state (just as, as a matter of fact, a severed maple tree branch is not a branch of a particular maple tree anymore). One example of this type of cases is *pretence*.

According to Bar-on, cases of pretence are cases in which someone expresses a certain mental state without expressing *her own* mental state (because she doesn't have such a mental state as a whole). Imagine someone who is pretending to have a headache to rid himself from cleaning the house, so he says with a certain tone, facial expression and bodily posture: "I have a terrible headache". The facial expressions, the bodily posture and the avowal "I have a terrible headache" are expressions of S that express a mental state of headache because there is a *conceptual connection* between those expressions and the mental state of headache: they are aspects or components that can possibly take place in a mental state of headache as a whole. In Bar-on's terms (Bar-on, 2004, p. 248), the facial expressions and the bodily posture Exp_1 (i.e., express in the action sense), and the avowal "I have a terrible headache" both Exp_1 (i.e., express in the action sense) and Exp_3 (i.e., express in the semantic sense), a mental state of headache. However, those expressions of S cannot express the headache *of* S because in this case, as a matter of fact, S doesn't have a headache (as a whole mental state). In Bar-on's terms, neither the avowal, the facial expressions nor the bodily posture Exp_2 (i.e., express in a causal sense) the mental state of headache of the subject because the subject doesn't have a headache (as a whole mental state), and so, S's expressions cannot be caused by a mental state of headache (they have to be caused by a different mental condition instead; e.g., by the intention to deceive)[16].

The idea that S's expressions are *defeasible criteria* of S's mental states follows from the account of pretence that we've just seen. Since expressions are considered to be *components*

[15] "A subject may display behavior expressive of pain (onstage, in pretense, to deceive, etc.) which, on the given occasion, does not express her pain." (Bar-on, 2004, p. 419).

[16] Notice that Bar-on uses here the concept of causality in a different way than the accounts that assume that expressions are *symptoms* of mental states (e.g., Armstrong, 1995, pp. 175-190; Putnam, 1975, Vol. 2, pp. 325-341; Shoemaker, 1996). These accounts think that mental states and expressions are different and independent items which are *externally* related by cause and effect. However, Bar-on thinks that between expressions and mental states there are *internal* causal relations (with both a conceptual and an empirical character) because expressions are *components* of mental states and the components of an item can typically establish causal relations among them. Thus, in Bar-on's view, mental states and expressions are considered to be *different* items, but they are not considered to be *independent* items.

of mental states, expressions and mental states are always and necessarily *conceptually connected* by the kind of thing that they are (i.e., fully-fledged mental states and possible aspects of those mental states). So, S's expressions are *criteria* of S's mental states. However, since it is still possible that some S's expressions take place without presenting an aspect of a mental state *of* S (because S doesn't have such a mental state as a whole), S's expressions are *defeasible* criteria of S's mental states. Therefore, it follows that between S's expressions and S's mental states there is an *internal relation* with both an *empirical* and a *conceptual* character. On the one hand, insofar as S's expressions are empirically related to S's mental states, it might still happen (i.e., when the criterion is *defeated*) that S's expressions take place without S's corresponding mental state taking place as well, just as a maple tree branch can take place without a maple tree because it has been *detached* from its trunk. For instance, to fake a headache saying "I have a terrible headache" expresses headache although not *my* headache (for I'm perfectly fine, and so, I don't have the mental state of headache as a whole). On the other hand, insofar as S's expressions are always and necessarily conceptually connected to a particular *type* of mental state, they are always and necessarily criteria (either defeated or not) of a possible mental state of S. Indeed, a branch severed from a maple tree is still a *maple tree* branch because it is a token of a *type of component* of maple trees (namely, a token of maple tree branches), and so, it is still conceptually connected to maple trees even if it is severed from the trunk and it is not anymore the maple tree branch of a particular maple tree. Analogously, S's expressions (e.g., to say "I have a terrible headache") are always *aspects* or *components* of a type of mental state that S could have (e.g., a whole headache), and so, they are always and necessarily conceptually connected to that particular type of mental state (being so always and necessarily criteria of that mental state).

Finally, from endorsing the idea that expressions are *indefeasible criteria* of mental states follows the non-relational view of expression endorsed by behavioural expressivism. According to this non-relational view, S's expressions and S's corresponding mental states are *a single set of items*, and so, mental states cannot be further items to which expressions are related in some way because mental states are nothing over and above a particular set of expressions. Inspired by Wittgenstein[17], the behavioural-expressivist account considers that

[17] Take, for instance, the following remark from the *Philosophical Investigations*:

> "'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.

mental states are *patterns of expressive behaviour* (e.g., smiling, crying, moaning, saying something, shouting something, doing something, having a certain bodily posture, etc.) manifested by a subject over a certain period of time (i.e., during the period of time that the subject has the particular mental state). So, mental states are *temporal processes*, dynamic realities that can only be appropriately studied in a diachronic way rather than fixed realities that can be appropriately studied in a synchronic way. As a result, mental states and expressions are not two different *items* but two different *perspectives* from which a single item can be considered. On the one hand, when a mental state is considered *diachronically*, it is seen as a *fully-fledged temporal event*, namely, as a whole pattern of expression extended over time (e.g., belief, hope, desire, happiness, pain…). On the other hand, when a mental state is considered *synchronically*, it is seen as an *aspect* of a pattern of expression, namely, as a particular expressive manifestation of a pattern that takes place in a particular moment (e.g., an action, a facial expression, an utterance, a smile, some tears, a bodily posture, a tone of voice, etc.). As a result, some bodily manifestations of a subject (e.g., a smile) have a particular *intrinsic expressive content* (e.g., happiness) because of the fact that they are *aspects* of the expressive pattern of a particular mental state (i.e., because of the kind of thing that they are) and not because of the fact that they are related to a particular mental state (for there is no mental state as a further item to which being related).

Therefore, while relational views of expression consider that having expressive content is a *relational property* (i.e., a property that some manifestations of S have because they are related to a further item or mental state), the behavioural-expressivist account claims that having expressive content is an *intrinsic property* (i.e., a property that some manifestations of S have because of the kind of thing that they are and not because of a relation to a further item). On the one hand, cause-and-effect relational views consider that expressions are the manifestations of S that have expressive content because there is an external causal relation between them and a particular mental state of S. As a result, cause-and-effect relational views of expression think that tears of allergy, *pretended* tears of sadness and actual tears of sadness are three manifestations of the subject that are *numerically different* and *typologically identical*: all of them are tears and only tears. However, while in the first case the tears don't express anything because they are not causally related to any mental state, in the second and third cases the tears express either the intention to pretend sadness or actual sadness because they are

'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).

causally related to the intention to pretend sadness or to actual sadness (which are conceived as further items). On the other hand, constitutive or mereological relational views consider that expressions are the manifestations of S that have expressive content because they are internally related to a mental state in the same way that a part is supposed to be internally related to the whole (when one endorses a defeasible notion of criteria): with an internal relation that has both a conceptual and an empirical character. Expressions are always conceptually related to a mental state because they are always possible aspects or components of a mental state; and expressions can be empirically related to a mental state because (as a matter of fact) they can take place within the totality of a mental state of S. As a result, constitutive or mereological relational views of expression think that tears of allergy are *numerically* and *typologically different* both from pretended tears of sadness and from tears of actual sadness, but that pretended tears of sadness and tears of actual sadness are *numerically different* and *typologically identical*. For only pretended tears of sadness and tears of actual sadness are an aspect of the world of the type that *can be* a component of the mental state of sadness, being so manifestations that have the expressive content of sadness because of the conceptual relation that they have to the mental state of sadness (which is conceived as a further item).

By contrast, the non-relational view of expression considers that having expressive content is an *intrinsic property* (i.e., a property that some manifestations of S have because of the kind of thing that they are and not because of a relation to a further item). As a result, the non-relational view of expression considers that tears of allergy, pretended tears of sadness and tears of actual sadness are all *numerically* and *typologically different*. Tears of allergy don't have expressive content, pretended tears of sadness have the expressive content of pretended sadness and tears of actual sadness have the expressive content of actual sadness. The expressive content, or lack of expressive content, of those three manifestations depends on the kind of things that they are and not on any relation to a mental state (for there is no mental state as a further item to which being related). Tears of allergy don't have expressive content because they are not an *actual aspect* of any mental state, pretended tears of sadness have the expressive content of pretending sadness because they are an *actual aspect* of the mental state of pretending sadness (more about this in section 2.1.3.), and tears of actual sadness have the expressive content of sadness because they are an *actual aspect* of the mental state of sadness.

Finally, it is relevant to point out that the behavioural-expressivist view of mental states has an advantage over Bar-on's mereological or constitute view of mental states. All the aspects or components of a three-dimensional object are there in the three-dimensional space even if it is not possible for the same subject to perceive them all at the same time: when some aspects are available to me in perception, other aspects are necessarily hidden from me, and vice versa (e.g., the façade and the interiors of a house cannot be perceived at the same time by the same subject). Analogously, in a mereological or constitutive view, it is not possible to perceive all the aspects or components of a mental state at the same time because only a few of them can be expressed or manifested by the subject at a given time (e.g., I might express happiness by smiling, but my smile is just an aspect or component of my mental state of happiness as a whole). Then, if mental states are supposed to be analogous to three-dimensional objects, when a subject has a particular mental state, *where* are supposed to be the aspects or components of that mental state that are currently hidden from other people's perception because they are not currently expressed or manifested by the subject? If Bar-on wants to give a complete account of mental states, she has to answer this question. And it seems difficult that she could answer this question without assuming ontological commitments that could go against the spirit of her neo-expressivist account (e.g., claiming that the aspects of the mental states that are currently unexpressed are in the nervous system of the subject). However, this problem doesn't arise to the behavioural-expressivist account insofar as it conceives mental states as temporal processes. Since the behavioural-expressivist account considers that mental states are patterns of expression extended over time, the expressions that form a mental state are considered to be *episodes* whose occurrence is scattered over the period of time that the subject is in that mental state (e.g., during the period of time that I am happy). So, the question of *where* the aspects or episodes of mental states that are *not currently* manifested or expressed by the subject are supposed to be doesn't arise: they are *currently* nowhere.

In summary, it has been argued in this chapter that from the idea that expressions are symptoms of mental states follows a cause-and-effect relational view of expression, from the idea that expressions are defeasible criteria of mental states follows Bar-on's mereological or constitutive view of expression, and from the idea that expressions are indefeasible criteria follows the non-relational behavioural-expressivist account of mental states. According to relational views, the expressive content of the behaviour of a subject is a relational property (i.e., a property that depends on a relation to a further item). By contrast, according to the non-relational view of expression, the expressive content of the behaviour of a subject is an intrinsic property (i.e., it depends on the kind of thing that the behaviour is). Also, it has been pointed out that the non-relational view of expression has the advantage of not having to explain where are the aspects of the mental state that are not being currently expressed by the subject because mental states have a temporal nature (i.e., they are progressively deployed over time). In the next section, the behavioural-expressivist account of expression and mental states is going to be explicated in detail.

2.1.2 The behavioural-expressivist view of mental states

In this section, the main ideas of the behavioural-expressivist conception of mental states are going to be developed in more detail[18]. The behavioural-expressivist account considers that mental states are nothing over and above expressive behaviour. Particularly, a mental state (e.g., belief, desire, intention, pain or happiness) is identical to a *pattern of expression* (Fig.1) manifested by the subject over a period of time (during which the subject has the particular mental state). For instance, if someone is sad because his football team was defeated in the final of the Champions League, he will manifest some scattered expressive behaviour of sadness during the period of time that he is sad: from the moment that he realized in the middle of the second half of the match that his football team was already defeated until he stops being sad because of that (maybe a week or two later). Each mental state or pattern of expressive behaviour is composed of a set of *episodes of expression* (Fig. 2) that are distributed in a certain way over the period of time that the expressive pattern lasts (i.e., during the period of time that the subject has the mental state). Thus, an episode of expression is a particular manifestation of an expressive pattern or mental state that occurs at a specific point (e.g., at t_2) of the time interval that the expressive pattern or mental state lasts ($T_{0\text{-}n}$). For instance, imagine that the subject from the last example cries out of sadness when the referee blows the whistle to end the match. This cry is an expressive episode of his mental state of sadness because it is an expression of sadness that takes place at a specific point in the time interval of his expressive pattern of sadness; particularly, when the referee blows the whistle to end the match.

Therefore, mental states and expressions are two different *perspectives* from which a single item can be considered. When the item is considered as a mental state (e.g., sadness), it is conceived as a whole pattern of expression manifested by the subject over a certain period

[18] The following explication of the non-relational view of expression takes the basic ideas from García (2018) and it tries to expand upon them to enrich the account.

of time (e.g., during the period of time that the subject is sad). And when the item is considered as an expression (e.g., crying out of sadness or saying a few days later "I'm still sad because of the football match"), it is conceived as a particular episode or manifestation that takes place at a specific point (e.g., at t_2) of the time interval that the mental state lasts ($T_{0\text{-}n}$). Before continuing with the explanation, it might be helpful to see a graphical representation of how the typical expressive pattern of a mental state might look like:

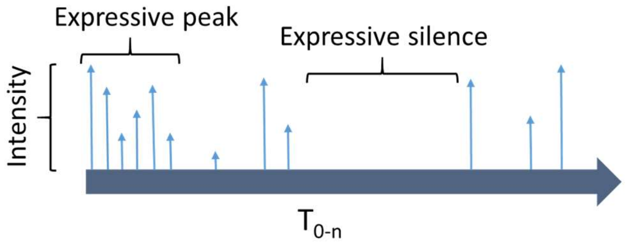

Fig. 1: Pattern of expression.

Fig. 2: Episode of expression.

A *vehicle of expression* is a piece of material or bodily behaviour (e.g., a smile, a cry, a moan, a frown, a bodily posture, a movement, some sounds coming from my vocal tract, etc.) that may occur in the *context* of the temporal expressive pattern of a mental state or may not. When a vehicle of expression doesn't occur in the context of the temporal expressive pattern of a mental state, it doesn't bear any expressive content (i.e., it is an empty vehicle of expression) because it is not an expressive episode of any mental state. For instance, a few tears (vehicle of expression) caused by an *allergy* don't express any mental state (i.e., they are an empty vehicle of expression) because they don't occur in the context of the temporal expressive pattern of any mental state. By contrast, when a vehicle of expression occurs in the context of the temporal expressive pattern of a mental state, it bears the (intrinsic) expressive content of the mental state in question because it is constitutive of an expressive episode of that mental state. For instance, when some tears (vehicle of expression) occur in the context of the temporal expressive pattern of sadness, they express sadness (i.e., they bear the expressive content of sadness) because they are constitutive of an expressive episode of sadness.

Therefore, a non-empty *vehicle of expression* (e.g., a smile) is an item that can be considered from two different conceptual perspectives: either as being a piece of material behaviour or as bearing a certain expressive content. When the vehicle of expression is seen as a piece of material behaviour, it is seen as a *material item* (e.g., a certain distribution of the flesh and muscles of the face of a subject) that occupies a place in a network of material causes and effects (e.g., that it is caused by a nerve impulse and that causes a certain reflection of the light that strikes the retina of the eyes of another subject). By contrast, when the vehicle of expression is seen as bearing an expressive content (e.g., happiness), it is seen as being constitutive of an *expressive episode* that occupies a place in the context of the temporal expressive pattern of a particular mental state; therefore, it is seen as a mental item that occupies a place in the network of mental causes and effects (e.g., she smiled because she saw the unexpected appearance of a good friend, who also smiled in response).

Two additional ideas need to be explained to further clarify the concepts of *expressive episode* and *expressive vehicle*. Firstly, a single expressive vehicle can have the expressive content of several mental states at once because it can take place in the context of several expressive patterns at the same time. Hence, a single expressive vehicle can be constitutive of expressive episodes of different mental states at once. For instance, my action of picking up the umbrella (vehicle of expression) is an episode of my belief that it is raining, an episode of my desire not to get wet and an episode of my intention to pick up the umbrella, having so my action the expressive content of my belief, of my desire and of my intention at once. For my action of picking up the umbrella (expressive vehicle) occupies a place in the context of the temporal expressive patterns of those three mental states at the same time. By contrast, there are other cases in which a single vehicle of expression is constitutive of a single expressive episode of mental state. For instance, a grimace (vehicle of expression) is an expressive episode of headache and only an expressive episode of headache when it takes place in the context of the temporal expressive pattern of headache and when it doesn't take place in the context of the temporal expressive pattern of any other mental state.

Secondly, an expressive episode of mental state can be formed of one vehicle of expression or of more than one vehicle of expression. For instance, to cry out of sadness is a single expressive episode of sadness (i.e., a manifestation of the temporal expressive pattern of sadness that takes place at a specific point in time) that is formed of multiple vehicles of expression at once: a characteristic facial expression, some tears, a characteristic demeanour, saying "We were so close this time!" with a sad tone of voice, and so on. All these vehicles of expression form a single expressive episode of sadness (i.e., a cry) when all of them occupy the *same place* in the context of the temporal expressive pattern of sadness (i.e., when all of them are manifested at the same point of the time interval of the expressive pattern of sadness). By contrast, to allow a subtle grimace of annoyance at what someone is saying is an expressive episode of annoyance that is formed by a single expressive vehicle: the grimace. The grimace alone is an expressive episode of annoyance because the grimace alone occupies a specific place in the context of the temporal expressive pattern of the mental state of annoyance (i.e., it occupies alone a specific point of the time interval of the expressive pattern of annoyance).

However, mental states don't take place in the void. Mental states are *ways of being in the world* because they are ways of interacting with other subjects and with other aspects of the world. Then, it is time to explain another aspect of the expressive patterns of our mental states: their *intentional object*. In the same way as a vehicle of expression (e.g., a smile) can be considered either from the perspective of being a piece of material behaviour (e.g., a certain distribution of the flesh and muscles of the face) or from the perspective of being an episode of expression (e.g., a smile of happiness), the *expressive content* of an episode of expression (e.g., happiness) can be seen from two conceptual perspectives as well: either as *presenting an aspect* of a mental state of the subject or as being *related to* a certain aspect of the world (in an appropriate or inappropriate way). When the expressive content is seen as being a presentation of a mental state of the subject, it is seen as an expressive episode of the pattern of a mental state of the subject. By contrast, when the expressive content is seen as being related to an aspect of the world (in an appropriate or inappropriate way), it is seen as having *intentionality*. For instance, imagine that I spontaneously smile when I see a good friend unexpectedly appearing at my house party. The expressive content of my smile can be considered from two different perspectives: either as being a *presentation* of my mental state of happiness (i.e., as an expressive episode of happiness) or as being *related* in a certain way to my friend (i.e., as having an intentional object: my friend). If the person at whom I am smiling is actually my friend, the appropriate relation of fit between my smile and its intentional object takes place. By contrast, if the person at whom I am smiling is not actually my friend but another person with the same haircut, the appropriate relation of fit between my smile and its intentional object doesn't take place.

In summary, the different conceptualizations of the non-relational view proposed here are as follows. Mental states are ways of being in the world or patterns of expressive behaviour. Therefore, mental states can be considered from two different perspectives: either as whole patterns of expressive behaviour or as an array of particular episodes of expression distributed in a certain way over time. In turn, an episode of expression is a piece of bodily behaviour or expressive vehicle that bears a certain expressive content because of the fact that it occurs in the context of a temporal pattern of expression (i.e., in the context of a mental state). That's why a non-empty vehicle of expression can be considered either from the perspective of being a material item or from the perspective of being an episode of expression. Finally, the expressive content of an episode of expression can be considered either as being an aspect or presentation of a mental state of the subject or as being related to an aspect of the world in an appropriate or inappropriate way (i.e., as having intentionality).

Let's illustrate these ideas with the example of grief. Grief is a good example to show that mental states are patterns of expressive behaviour because it is a mental state that evolves over time through the different phases pointed out by psychology. That's why it has been used by philosophers on other occasions (e.g., Goldie, 2011; Wittgenstein, 1953). According to the behavioural-expressivist account, the mental state of grief is nothing over and above a certain way of being in the world, namely, a pattern of expression with its intentional object. So, the expressive pattern of grief is 1) a selection of *expressive episodes* characteristic of grief, 2) distributed over the *time period* that is characteristic of grief and 3) distributed *in the way* that is characteristic of grief over that time period. Also, all the expressive episodes of this pattern have the same intentional object: the beloved person who passed away.

Firstly, 1) the set of episodes of expression that we consider characteristic of a grieving subject includes both *linguistic* and *non-linguistic episodes*. Among the non-linguistic episodes of expression that a grieving subject might manifest are: crying, having a sad facial expression and bodily posture, being negative, having a bad mood (i.e., getting easily angry or annoyed), doing things in a lazy way, seeking to stay alone for more time than usual, being apathetic when forced to socialize by the situation, etc. And among the linguistic episodes of expression that a grieving subject might manifest are: saying how much she misses the person who passed away, remembering happy moments that they spent together in the past, telling anecdotes about him, etc. All these episodes of expression might vary in *intensity* (Fig. 1). Indeed, it is not the same to say "I still miss him" with a restrained tone of voice as to inconsolably shout "I still miss him!"; or it is not the same to shed a tear while talking about him as it is to break down in tears. Moreover, not only expressive episodes have *intensity* but also the expressive patterns themselves. For it is clear that it is possible to grieve a person more or less (i.e., with different intensity). The intensity of a mental state depends on both the amount and the intensity of the episodes of expression that compose its expressive pattern. The higher the number of expressive episodes and the higher their intensity, the more intense a mental state will be (as long as the pattern of the mental state is not blurred or faded out by excess into the expressive pattern of something else).

Secondly, 2) the *temporal extension* of the expressive pattern ($T_{0\text{-}n}$) is also a constitutive element of the expressive pattern itself (i.e., of the mental state). There are different periods of time characteristic of different types of expressive patterns or mental states. For instance, the temporal extension of the expressive pattern of grief is somewhat comprehended between some days (at least) and some years (at most)[19]. On the one hand, if someone supposedly grieved a person for only 10 minutes, we would strongly doubt her sincerity, since it is not possible to genuinely grieve a person over that very short period of time. On the other hand, if someone supposedly grieved a person for more than 10 years, we would tend to think that he is depressed or nostalgic rather than grieving, since it is not possible to *genuinely* grieve a person for more than 10 years. By contrast, there isn't any problem with the possibility of someone having a headache for 10 minutes or with the possibility of someone having the belief that the Earth goes around the Sun for more than 10 years. Therefore, the temporal extension ($T_{0\text{-}n}$) of the expressive pattern of a mental state is also a constitutive element of the expressive pattern itself.

Finally, 3) it is constitutive of the expressive pattern of grief that there are moments of *expressive peaks* and moments of *expressive silences* (Fig. 1). Expressive peaks are periods of time in which the subject manifests a higher number of expressive episodes, and of higher intensity, of a particular mental state. Expressive peaks might be triggered by a certain situation of the life of the subject. For instance, a grieving person might manifest a peak of expressive episodes (i.e., crying, saying "I miss you" in a *sotto voce*, remembering past experiences, etc.) if she is alone and she finds a picture of the beloved person who passed away. By contrast, expressive silences are periods of time in which the subject doesn't manifest any expressive episodes of a particular mental state. Expressive silences might be favoured by situations that are neutral towards that particular mental state or by situations that don't require the manifestation of that mental state. For instance, someone who is grieving a person for months will have periods of time in which she doesn't manifest any episode of grief because she is having a moment of enjoyment focused on other things, because she feels better that day for some reason, because she is sleeping, etc.

However, not only is it constitutive of the mental state of grief that there are moments of expressive peaks and moments of expressive silences, but also that there is a certain *temporal distribution* of those episodes of expressive peaks and expressive silences. Firstly, it is clear

[19] This analysis is meant as an example to argue that the temporal extension of the expressive pattern of a mental state is a constitutive element of the expressive pattern itself. So, it is important to clarify that the precise temporal extension of grief, as well as the precise temporal extension of any other mental state, is a matter of empirical investigation (and not of conceptual investigation).

that it belongs to the natural course of grief that expressive peaks are more frequent and intense at the beginning (when the loss of the beloved person is recent), and that as time goes on, they become gradually less intense and gradually replaced by expressive silences (until the subject finally overcomes her loss and stops grieving). Secondly, it is clear that expressive peaks and expressive silences of grief have a limit *by excess* and a limit *by defect*. On the one hand, if a subject manifested expressive peaks of alleged grief *all the time* for a long period of time (days, weeks…), the expressive pattern implemented by the subject wouldn't be one of grief but one of a different mental state (e.g., a mental breakdown because of the loss of a beloved person), and so, it would be appropriate to attribute that mental state to her instead. On the other hand, if a subject didn't manifest any expressive episode of grief at all (as if she were in an alleged constant expressive silence), no expressive pattern of grief would have been implemented by the subject, and so, it wouldn't be appropriate to attribute the mental state of grief to her. Therefore, both expressive peaks and expressive silences, as well as a certain temporal distribution between them, are constitutive elements of grief.

It is important to notice that from the fact that both expressive peaks and expressive silences are constitutive elements of grief (as well as of *most* mental states), it doesn't follow that they are constitutive elements of *all* mental states. In fact, there are mental states whose pattern of expression doesn't have any expressive silence. Imagine that I am alone in my office when I hit my knee with the desk, having a strong pain that lasts about 15 seconds. Imagine that, as a result, I moan out of pain, I curse the desk and I get up from the chair just to limp around the room with a grimace of pain until the pain ceases. Then, in this case, I have a mental state of pain whose expressive pattern (that lasted for about 15 seconds) has one expressive peak and no expressive silence at all.

So far I have described the *typical* expressive pattern of grief to characterize the behavioural-expressivist notion of mental state. However, the typical expressive pattern of a mental state is an *abstraction* or *generalization* made from the different *instances* of expressive patterns of that mental state that are actually implemented by particular subjects. Mental states don't take place in the void but in the context of a *form of life* (i.e., in the context of the *interactions* between the subject who has the mental state, other individuals of her community and the world; interactions aimed at performing a variety of activities), and so, the *typical* expressive pattern of a mental state (e.g., grief) might be *instantiated* in multiple ways by different subjects depending on 1) *the idiosyncrasy and cultural background* of the subject and 2) the stimuli provided by her *context and day-to-day situations*.

Firstly, 1) the way in which the typical expressive pattern of a mental state is instantiated by a subject is affected both by the idiosyncrasy of the subject and by her cultural background. For the *episodes of expression* that the subject will tend to use to express a certain mental state (e.g., some subjects might be prone to express their sadness by crying, while others might be prone to express their sadness by talking) and the *vehicles of expression* that the subject will tend to use to produce those episodes of expression (e.g., there are different ways to cry, different ways to laugh, different ways of talking, etc.) depend on the socialization process of the individual within her community. Indeed, there are *natural* episodes of expression that are characteristic of human beings and other animals because of the kind of natural beings that they are (e.g., crying out of pain or crying out of hunger). But, during the socialization process of the individual, some of those natural expressions are tamed, and sometimes replaced, by *culturally learnt* episodes of expression (e.g., we are trained to say "It hurts" or "I'm hungry" instead of crying out of pain or hungriness). As a result, differences between the actual way in which different subjects instantiate an expressive pattern arises. Differences both regarding the *type of expressive episodes* that they manifest (e.g., crying vs. saying "I still miss him") and regarding the *expressive vehicles* that they use to produce those expressive episodes (e.g. crying or talking in one way or another). Furthermore, not only among different individuals there are differences but also among different communities. For instance, one difference between the way in which the expressive pattern of pain is instantiated in an English-speaking community and in a Spanish-speaking community is that in the former the expressive vehicle "It hurts" will likely be an episode of pain, while in the latter the same expressive role will be likely played by the expressive vehicle "Me duele".

Secondly, 2) the way in which the expressive pattern of a mental state is instantiated by a particular subject (i.e., the description that her instantiation of the expressive pattern will have) depends also on the day-to-day contexts and situations of the subject because the expressive episodes of a pattern of expression have a *dispositional character*: their actual manifestation at a particular time (e.g., t_3) depends on the stimulus provided by the current situation of the subject. Indeed, it is a condition of having a mental state during a certain period of time ($T_{0\text{-}n}$) that the subject *actually* manifests episodes of such mental state over $T_{0\text{-}n}$ (at least a single episode). For otherwise, without the *actual* occurrence of any expressive episode of that mental state, there wouldn't be any expressive pattern of that mental state, and so, the subject wouldn't have the mental state. However, even if some *actual* episodes of expression have to occur over $T_{0\text{-}n}$ to have a pattern of expression or mental state, expressive episodes have a dispositional character because their actual occurrence depends on the situation in which the subject is. For instance, a grieving subject has the disposition to manifest different expressive episodes of grief. These dispositions are actualized or not depending on the situation of the subject. If she saw a picture of the person who passed away or if a close friend asked her how she feels, she would likely manifest some expressive episodes of grief (i.e., crying, saying how much she misses him, etc.); but if she were distracted doing a task that requires her full attention or if she were sleeping, she wouldn't likely manifest any expressive episode of grief at all. Therefore, the expressive pattern of a mental state is instantiated in different ways depending also on the context and the day-to-day situations of the subject.

It is important to notice that the fact that there are a variety of ways in which the expressive pattern of a mental state can be instantiated by different subjects (depending on the subject's idiosyncrasy, community and day-to-day situations) doesn't mean that we are unable to recognize the *type* of mental state that a particular instance of expressive pattern is. For in spite of the differences between the different instances of expressive patterns, there are a few elements that allow us to recognize a particular instance of expressive pattern as belonging to a certain type of mental state (i.e., belief, desire, pain, happiness, grief, enjoyment, etc.). Firstly, insofar as expressive patterns have a natural basis (i.e., there are natural expressions on the basis of each one of our idiosyncratic expressive patterns), there is a *family resemblance* among the particular ways in which different individuals instantiate a particular type of expressive pattern (i.e., a particular type of mental state). Secondly, even if every individual has an idiosyncratic way to instantiate a particular expressive pattern (e.g.: a way of being sad, a way of being in pain, a way of being happy, etc) and its particular expressive episodes (e.g., a way of crying out of sadness, a way of moaning out of pain, a way of laughing out of happiness, etc.), it is still possible to get to know the individual in question and her particular ways of expression. And thirdly, since patterns of expression have a temporal nature, it is possible to observe the subject for some period of time to put her expressive episodes into the contexts of a particular type of expressive pattern. All these are elements that allow us to properly identify the *expressive content* of the expressive episodes of a particular subject (e.g., whether this smile is an episode of the pattern of happiness or of the pattern of nervousness), overcoming so the differences between instances of expressive patterns and identifying the *type* that a particular instance of expressive pattern belongs to.

2.1.3 Pretension and dissimulation: two cases for behavioural expressivism

Relational views of expression could argue against the non-relational view of expression endorsed by the behavioural-expressivist account that only relational views can explain why subjects can *pretend* that they have a mental state that they don't actually have or *dissimulate* that they have a mental state that they actually have. For relational views could argue that only claiming that mental states and expressions are two different sets of items that can take place separately in a subject it is possible to explain pretence and dissimulation. For instance, Bar-on (2004, pp. 418-420) argues that subjects can pretend to have a mental state that they don't have because they can produce the characteristic set of expressions of that mental state without having the mental state and that subjects can dissimulate that they are currently having an episode of a particular mental state because they can actively suppress the characteristic expressions of that mental state. However, not only the non-relational view of expression endorsed by the behavioural-expressivist account can actually explain why subjects can pretend and dissimulate mental states, but it has also been argued (García, 2018) that it explains pretence and dissimulation better than relational views because it explains them with less theoretical resources. Let's see first how the behavioural-expressivist account explains pretence and dissimulation, and then, why it explains them better than relational views of expression.

According to the behavioural-expressivist account, subjects can pretend that they have a mental state that they don't actually have because *pretending to have* the mental state M (e.g., pretending to have a toothache) is a *sui generis* mental state, different from the mental state M (e.g., having a real toothache), that *mimics* some of the expressive episodes of the mental state M (e.g., some of the expressive episodes of a real toothache). Indeed, a person who pretends to have M will manifest a pattern of expression as similar in appearance as possible to some of the aspects of the expressive pattern of having M. However, similar doesn't mean identical; otherwise, we would be talking of one expressive pattern rather than two. The expressive pattern of pretending M and the expressive pattern of having M are *different* expressive patterns because they are different in regard to three aspects. The first aspect is related to the *temporal distribution* of the expressive episodes; for the expressive episodes of the pattern of pretending M and of the pattern of having M are usually manifested in different situations. The second aspect is related to the *way* in which the expressive episodes of each mental state are usually manifested; for even if some expressive episodes of the pattern of pretending M mimic some expressive episodes of the pattern of having M, there are usually subtle differences between the expressive episodes of having M and their counterfeits. And the third aspect is related to the *number* of expressive episodes that the pattern of pretending M and the pattern of having M have; for the pattern of pretending M has (at least) an expressive episode that has no counterpart in the pattern of having M: the expressive episode of *confessing* that one is pretending to have M (remember that the episodes of an expressive pattern are dispositional so that they are not necessarily manifested in every particular case).

Imagine a subject who is pretending to have a toothache to rid himself from cleaning the house. Some of the expressive episodes of the pattern of a subject who pretends to have a toothache will *mimic* some of the expressive episodes of the pattern of a subject who has a real toothache by mimicking their vehicles of expression (e.g., moaning, having a particular facial expression, closing her eyes, touching her cheek, saying "I have a terrible toothache"…). However, in spite of that, both patterns will be different in regard to the three aspects mentioned before. Firstly, the *temporal distribution* of the expressive episodes of each mental state will be different because they will tend to be actualized in different situations. For instance, the subject who is pretending to have a toothache will stop pretending (e.g., moaning, having a particular facial expression, closing her eyes, touching her cheek, taking an aspirin, saying "I have a terrible toothache"…) when she thinks that she is alone in the house and that nobody is watching, while the subject who has a real toothache will express episodes of toothache (e.g., moaning, having a particular facial expression, closing the eyes, touching her cheek, taking an aspirin, murmuring "I have a terrible toothache"…) even if she is alone and she thinks that nobody is watching. Secondly, even if some of the expressive episodes of a subject who pretends to have a toothache and of a subject who has a real toothache are similar in appearance, there are normally *subtle differences* between them. The moans, facial expressions, gestures, actions, glances, utterances… of a subject who pretends to have a toothache and the moans, facial expressions, gestures, actions, glances, utterances… of a subject who has a real toothache are usually different in their details. And thirdly, the expressive pattern of pretending to have a toothache has (at least) an expressive episode which has no counterpart in the expressive pattern of having a toothache: the *confession* that one is pretending to have a toothache to rid himself from cleaning the house. Therefore, the expressive episodes of each of these mental states have different expressive contents because they belong to different patterns of expression (i.e., to different mental states). The expressive episodes of a subject who pretends to have a toothache will have the expressive content of *pretended toothache* and the expressive episodes of a subject who has a real toothache will have the expressive content of *real toothache*.

From these three differences in the expressive patterns of pretending to have M and of having M follows the idea that the *perfect pretender* cannot exist due to conceptual reasons: a pretender can always be discovered by the appropriate observer. It is always *possible* for a person who knows the pretender well and who spends enough time with her to tell whether she is pretending to have M or whether she actually has M because of the following three facts. Firstly, the possible existence of certain distinguishing details in the expressive episodes of the subject (e.g., in the way she moans, in the way she says "I have a toothache", in the way she closes the eyes and touch her cheek, etc.). Secondly, the revealing temporal distribution of the expressive episodes of the subject across different situations or contexts (e.g., the observer might catch the pretender enjoying a cold soda with a happy face when she thinks that she is alone and that nobody is watching, something that a subject with a strong toothache wouldn't be able to do). And thirdly, the subject could always end up confessing that she is pretending to have a strong toothache to rid herself from cleaning the house.

Granted, sometimes the act of pretending might be successful and the observer might be misled into thinking that the subject actually has the mental state M (even if she knows the pretender very well and she spends a lot of time with her). However, what that means is that the observer has mistakenly taken the expressive episodes of the subject as belonging to the mental state of M instead of as belonging to the mental state of pretending M. Thus, the success or failure of an act of pretension depends on two factors. On the one hand, it depends on the *epistemic competence* of the observer to tell the difference between the two kinds of expressive patterns (i.e., pretending M and having M). On the other hand, it depends on the *ability of the pretender* to mimic some of the expressive episodes of the pattern of having M; i.e., on whether the pretender manages to manifest episodes of pretending M more similar or less similar in appearance to the real expressive episodes of having M. If the pretender manages to manifest more similar episodes of expression, she will be a good pretender and her chances of success will increase; by contrast, if she doesn't manage to do so, she will be a bad pretender and her chances of success will decrease. Anyway, she will manifest expressive episodes of pretending to have M and not expressive episodes of having M (for she doesn't have M).

Finally, regarding *hiding* or *dissimulation*, subjects can hide or dissimulate that they are currently having an episode of the mental state M by suppressing the clearer and more intense expressive episodes of the mental state M. Imagine a subject who has a strong episode of headache at her birthday party but she hides or dissimulates that she has a headache in order not to ruin the event. This person might have the *disposition* to express her headache by manifesting clear and intense expressive episodes of headache (e.g., closing her eyes, making an obvious grimace of pain, saying "What a terrible headache I have!", etc.), but she will refrain from actualizing those clear and intense episodes of expression if she wants to hide or dissimulate her headache. However, even if a subject can suppress *some* expressive episodes of the mental state M, it follows from the behavioural expressivist-account that it is conceptually impossible for a subject to suppress *all* the expressive episodes of the mental state M while still having the mental state M. For, in that case, the subject wouldn't have the expressive pattern characteristic of M (which needs one expressive episode at least), and so, the subject wouldn't qualify as having the mental state M. Then, continuing with the example, every subject who hides or dissimulates that she has an episode of headache has to manifest at least one, and usually some, subtle expressive episodes of headache if it is true that she has a headache. For instance, she might make a subtle grimace of pain for a moment, she might answer a friend's question in a grumpy way because of her pain, or she might go to the toilet just to express her headache freely for some seconds.

Furthermore, in some cases, a subject can hide or dissimulate that she is having an episode of the mental state M not only by suppressing some of the most intense or clear episodes of M but also by *pretending* that she has a mental state that is normally incompatible with M. For instance, someone might hide or dissimulate that she has a headache by pretending that she is *enjoying* the party. So, she might smile in a forced way, she might try to talk with other people as if she was happy to be there, she might (insincerely) say with a forced tone of voice "How much I'm enjoying this party!", etc.

Therefore, the behavioural-expressivist account can explain both pretence and dissimulation even if it endorses a non-relational view of expression. Moreover, it has been argued (García, 2018) that the behavioural-expressivist account explains pretence and dissimulation better than relational views of expression. The reason why the non-relational view of expression offers a better explanation of pretence and dissimulation than relational views of expression is that it explains how people can *discover* that someone is pretending or dissimulating with the same few theoretical resources that it explains how people can have access to someone's mental states in normal cases (i.e., in cases which are not of pretence or dissimulation). On the one hand, relational views of expression explain dissimulation by claiming that subjects can have an episode of the mental state M without expressing the mental state M; and pretence by claiming that subjects can *express* the mental state M (i.e., neo-expressivist accounts), or produce the same manifestations as the mental state M (i.e., causal views), without having the mental state M. Then, from the perspective of the relational view of expression, the following question arises: how are people supposed to *discover* when someone else is pretending or dissimulating a mental state if to discover that involves *surpassing* her expressions to check whether she has the relevant mental state (i.e., dissimulation) or not (i.e., pretence)? Even assuming for the sake of argument that relational views of expression have an appropriate answer for this question, they need to introduce new theoretical elements into their accounts in order to explain how people are supposed to be able to check whether someone has the relevant mental state or not independently of her expressions; i.e., how people are supposed to be able to surpass someone's expressions or lack of expressions to check whether she has the relevant mental state or not.

On the one hand, causal views of expression consider that we don't have perceptual access to other people's mental states neither in normal cases nor in cases of pretence or dissimulation. Thus, causal views have to introduce new theoretical elements to explain how subjects can surpass other people's behaviour to find out which are their mental states both in normal cases and in cases of pretence or dissimulation. On the other hand, neo-expressivist accounts consider that people can have perceptual access to other people's mental states through their expressions. Thus, even granting that neo-expressivist accounts don't have to introduce new theoretical elements to explain how subjects can find out which are the mental states of other people in normal cases (for those mental states are supposed to be available to perception), they have to introduce new theoretical elements to explain how, in cases of pretence and dissimulation, subjects can *surpass* other people's expressions of M (when the subject *pretends* M) or lack of expressions (when the subject *dissimulates* M) in order to discover that they are pretending or dissimulating.

By contrast, the behavioural-expressivist account explains how people can discover pretence and dissimulation using exactly the same few theoretical resources that it uses to explain how people have access to someone's mental states in normal cases (i.e., in cases which are not of pretence or dissimulation). For pretending is all about manifesting a *sui generis* expressive pattern with episodes as similar as possible in appearance to the episodes of the pattern of the mental state that one is pretending to have, and dissimulating is all about suppressing the clearer and more intense expressive episodes of the pattern of the mental state that one is hiding. Then, to discover that someone is pretending or dissimulating a mental state is all about being able to *identify* the proper expressive content of the episodes manifested by the subject (i.e., to identify the proper pattern of expression) without being taken in by appearances, just as people do to have access to someone's mental states in normal cases (i.e., in cases which are not cases of pretence or dissimulation). As a result, the behavioural-expressivist account offers a better explanation of pretence and dissimulation, and of their discovery and identification by people, than relational views of expression. For it explains all these phenomena with the same theoretical resources, and so, using less theoretical resources than relational views of expression.

In this first half of the chapter, the implications between the way in which it is understood that expressions are evidence of mental states and the relational and non-relational views of expression have been explained first; afterwards, the non-relational view of expression has been explicated in detail, and it has been argued that the non-relational view of expression is preferable over relational views of expression because it explains how we can know other people's mental states, even in cases of pretence and dissimulation, with less theoretical resources than relational views of expression. In the remainder of the chapter, it is going to be argued that the idea of first-person self-knowledge as epistemic self-knowledge or true warranted belief (which is shared by neo-expressivist accounts and by epistemic accounts of Transparency) is conceptually flawed. As an alternative, a behavioural-expressivist account of first-person expressive self-knowledge and of third-person epistemic self-knowledge is going to be proposed.

2.2 First-person and third-person self-knowledge

There are two senses of knowing: *knowing that* and *knowing how* (Ryle, 1949). On the one hand, *knowing that* is an epistemic sense of knowing and it consists in having a true warranted belief. For instance, one has knowledge in the *knowing that* sense when one forms the true belief that the Earth goes around the Sun on the basis of appropriate evidence (e.g., a primary school book). On the other hand, *knowing how* is a non-epistemic sense of knowing and it consists in exercising an activity in a certain way. In turn, there are two senses of *knowing how*: one is a matter of *having the ability* to exercise an activity in an appropriate way (when the circumstances are adequate to exercise that ability), and the other is a matter of exercising that activity *self-consciously* as a matter of fact, namely, knowing what one is up to when performing the different steps of the activity. Since one cannot perform an activity knowing what one is up to (i.e., self-consciously) if one doesn't already have the ability to appropriately perform that activity in adequate circumstances, *knowing how* in the sense of having an ability is a condition of *knowing how* in the sense of performing that activity self-consciously. For instance, when one is able to swim appropriately in adequate circumstances (e.g., one is not drunk, the water is calm, etc.), one has knowledge in the *knowing how* sense that one has the *ability* to swim. And when one has the ability to swim and one exercises that ability knowing what one is up to (e.g., paying attention to what one is doing, to one's own swimming speed, to the distance remaining to finish the lap, to how tired one feels, to how much faster one can go, etc.), one has knowledge in the *knowing how* sense that one exercises an activity *self-consciously*[20].

By contrast, one can fail to have knowledge in the *knowing that* sense and in the *knowing how* sense in the following ways. On the one hand, one fails at having knowledge in the sense of *knowing that* when one ends up with a false or unwarranted belief. For instance, one fails to have knowledge in the *knowing that* sense if one ends up forming the false belief that the Sun goes around the Earth or if one ends up forming the true belief that the Earth goes around the Sun but the belief is unwarranted (e.g., it was formed on the basis of what I read in a magazine about pseudoscience). On the other hand, one fails at having knowledge in the *knowing how* sense when one doesn't have the ability to exercise an activity or when one exercises an activity *un-self-conscious*ly. Firstly, one doesn't have knowledge in the *knowing how* sense that one has an ability when one is not able to exercise an activity in an appropriate

[20] It could be objected that it is characteristic of people who master the ability to swim that they are able to swim without paying attention to the different steps of the activity (e.g., to the movements that they are making, to whether they are making them correctly, to the distance remaining to finish the lap…), and so, that the idea that performing an activity self-consciously is a kind of *knowing how* is mistaken (for it is clear that professional swimmers know how to swim and that they swim without paying attention to what they are doing). However, this objection is mistaken. Thanks to the fact that professional swimmers don't pay attention to the things that were pointed out before, they can pay attention to other steps of the activity that novel swimmers are not able to pay attention most of the time (e.g., whether they are going first in the race or whether they are able to swim a bit faster without getting too much tired before finishing the race). So, it is still true that professional swimmers don't swim *un*selfconsciously when they are competing (e.g., they are not thinking about what they are going to have for dinner), and so, it is still true that exercising an activity self-consciously is a kind of *knowing how*.

way even in adequate circumstances. For instance, when one is unable to swim appropriately even in circumstances that are adequate for the exercise of that activity (e.g., one is not drunk, the water is calm, etc.). Secondly, one doesn't have knowledge in the *knowing how* sense of self-consciousness when one exercises that ability on the given occasion without knowing what one is doing (i.e., un-self-consciously). For instance, when one swims without knowing what one is doing, either because one is swimming distractedly and without paying attention (e.g., thinking about what one is going to have for dinner) or because one doesn't have the ability to swim (e.g., one has just started to take classes of swimming).

Neo-expressivist accounts and the behavioural-expressivist account differ in their view of first-person self-knowledge. Neo-expressivist accounts consider, like epistemic accounts of Transparency, that first-person self-knowledge is an epistemic phenomenon involving a true warranted belief about one's own mental states (*knowing that*). For neo-expressivist accounts consider that first-person avowals are *self-ascriptions of mental states* that express the same mental state that they self-ascribed, and so, they argue that first-person avowals can give rise to true warranted beliefs about one's own mental states. By contrast, according to the behavioural-expressivist account of Transparency, the question "Do you believe that p?" is answered with a *judgement about p* when it is answered from the first-person deliberative perspective and with a *self-ascription of attitude* when it is answered from the third-person self-inspective perspective. As a result, the behavioural-expressivist account considers that first-person self-knowledge is a matter of the way in which subjects express their own mental states (*knowing how*) from the first-person deliberative perspective and that self-knowledge as an epistemic phenomenon (i.e., true warranted belief about one's own mental states) is a matter of self-inspecting oneself from the third-person perspective.

In the remainder of this section, it is going to be argued that the idea of first-person epistemic self-knowledge is conceptually flawed, and so, that first-person self-knowledge is a matter of *knowing how* because it is a matter of the way in which subjects express their mental states. Also, it is going to be argued that epistemic self-knowledge is an epistemic phenomenon resulting from the third-person process of self-inspection and that it is *sometimes* possible to acquire a *strongly warranted* second-order belief (in the sense that it is warranted to a *higher degree* and not in the sense that it enjoys an exclusive *type* of warrant) from the third-person perspective of self-inspection. Let's start by explicating the neo-expressivist account of first-person epistemic self-knowledge in detail.

2.2.1 The neo-expressivist account of first-person epistemic self-knowledge

Neo-expressivist accounts consider that first-person avowals (e.g., "I'm thirsty", "I believe that p" or "I want p") are *truth-evaluable* self-ascriptions of mental states that *express* the very same mental state that they self-ascribe. Firstly, they are *truth-evaluable* self-ascriptions of mental states because the avowing subject might have or might not have the self-ascribed mental state. When the avowing subject has the self-ascribed mental state, the avowal is true because the appropriate relation of fit between the self-ascription and the subject's mental state takes place. When the avowing subject doesn't have the self-ascribed mental state, the avowal is false because the appropriate relation of fit between the self-ascription and the subject's mental state doesn't take place. Secondly, first-person avowals are considered to be *expressions* of the self-ascribed mental state itself because first-person avowals are the result of the subjects *speaking up their minds* on the basis of no specific evidence about their mental states. So, first-person avowals are in a continuum with other non-truth-evaluable expressions (such as, linguistic interjections —e.g., "Ouch!", "Oops!" or "Hey!"— and natural expressions —e.g., tears of pity, smiles of happiness or moans of pain—) spontaneously manifested by subjects on the basis of no specific evidence about their own mental states.

As a result, neo-expressivist accounts consider that first-person avowals are *groundless* and *authoritative* first-person self-ascriptions of mental states (Bar-on, 2004, pp. 2-6; Finkelstein, 2003, pp.100-114). On the one hand, 1) first-person avowals are groundless self-ascriptions of mental states because they are expressions of the self-ascribed mental state (like a grimace of pain or the interjection "Ouch!") issued on the basis of no specific evidence about what is considered to be their subject matter (i.e., the subject's mental states). For if evidence were involved in the utterance of a first-person avowal, the evidence would be about the world (as required by Transparency) and not about whether one has the self-ascribed mental state (which is considered to be the subject matter of the self-ascription). On the other hand, 2) first-person avowals are considered to be authoritative, in the sense of enjoying a strong presumption of truth, because they are supposed to express the very same mental state that is the truth-maker of the self-ascription. So, denying the truth of a first-person avowal (i.e., denying that the subject has the self-ascribed mental state) would involve denying that the first-person avowal expresses the self-ascribed mental state in the way that it is supposed to do.

For instance, the first-person avowal “I believe that p” is supposed to express the same mental state that it self-ascribes (i.e., the belief that p). So, 1) it is supposed to be made on the basis of no evidence about what is considered to be its subject matter (i.e., whether I believe that p) because it is an expression of the belief that p itself and 2) it is supposed to enjoy authority or a strong presumption of truth because its truth-maker (i.e., my belief that p) is expressed by the avowal itself. It is true that the authority or presumption of truth of first-person avowals can be revoked if it is thought that there is some unusual circumstance (e.g., insincerity, unconscious mental states, self-deception, etc.) that prevents the first-person avowal from expressing the self-ascribed mental state (its truth-maker) as it is supposed to do; but in the absence of such unusual circumstances, first-person avowals are considered to be authoritative and to enjoy a strong presumption of truth.

By contrast, third-person avowals (which are avowals issued by self-inspection, like when Tom says: “I don’t actually believe in gender equality”), assertions about the world (e.g., “It is raining”) and assertions about other people’s mental states (e.g., “Lydia believes that it is raining”) are not supposed to be groundless or authoritative at all. In all these cases, 1) the assertion or self-ascription is supposed to be made on the basis of evidence about its subject matter (i.e., it is *not* groundless) and 2) the truth-maker of the assertion is supposed to be something different from the mental state expressed by the subject’s utterance itself. So, denying the truth of the assertion or self-ascription doesn’t involve denying that the subject has the mental state expressed by the utterance itself (and so, it is not authoritative because it doesn’t enjoy a strong presumption of truth). For example, the assertions “It is raining” and “Lydia believes that it is raining” express the *belief* that it is raining and the *belief* that Lydia believes that it is raining (respectively), but their truth-makers are something different: whether it is *actually* raining and whether Lydia *actually* believes that it is raining (respectively). Also, the third-person avowal “I don’t believe in gender equality” expresses the *second-order belief* that I don’t believe in gender equality, but its truth-maker is the different psychological fact of whether I have a *first-order belief* in gender equality or not.

Then, on the basis of the fact that first-person avowals are considered to be self-ascriptions of mental states that express the self-ascribed mental state (what is supposed to explain why first-person avowals are *groundless and authoritative* self-ascriptions), neo-expressivist accounts try to explain *first-person self-knowledge* (Finkelstein, 2003, pp. 148-152; Bar-on, 2004, Ch. 9). Neo-expressivist accounts understand first-person self-knowledge as an epistemic phenomenon (*knowing that*). So, in order to explain first-person self-knowledge, they need to show that their understanding of first-person avowals explains how subjects are able to acquire a *true strongly warranted second-order belief* when they issue a first-person avowal (e.g., "I believe that it is raining", "I want a piece of cake", "I have a headache", and so on). Particularly, they need to explain how the *true belief* involved in first-person epistemic self-knowledge is supposed to be about the subject's mental states and how the *warrant* of that belief is supposed to be *stronger* than the warrant of the beliefs involved in other kinds of knowledge. Otherwise, their account of first-person epistemic self-knowledge wouldn't be compatible with the fact that first-person avowals are groundless (i.e., based on no specific evidence about what is considered to be its subject matter) and authoritative (i.e., presumably true) self-ascriptions of mental states that express the self-ascribed mental state itself. Bar-on argues in detail that her neo-expressivist account is able to explain how a subject who avows a mental state from the first-person perspective is in the appropriate epistemic position to acquire a true strongly warranted belief about whether she has the self-ascribed mental state. Let's explicate Bar-on's account of first-person epistemic self-knowledge in order.

Firstly, to explain how a subject who avows a mental state from the first-person perspective can have a *belief* (i.e., a second-order belief) about whether she has the self-ascribed mental state, Bar-on argues in favour of what she calls the "dual expression thesis". Bar-on distinguishes two senses in which subjects can have a belief: the *opining sense* and the *holding-true sense*. On the one hand, to believe something in the opining sense is to judge that p is the case on the basis of some evidence while entertaining the thought that p. For instance, I believe in the opining sense that the Earth goes around the Sun if I have entertained that thought to judge, on the basis of certain evidence (e.g., that I read so in a science book), that it is true that the Earth goes around the Sun. On the other hand, to believe something in the holding-true sense is just to be disposed to assent to the truth of a certain content, on the basis of no specific evidence, if I were asked about that content (so that the subject *currently* has the non-manifested disposition to hold that truth). For instance, I believe in the holding-true sense that it rains in Mexico because if I were asked whether it rains in Mexico, I would immediately answer affirmatively on the basis of no specific evidence, even though I've never considered whether it rains in Mexico before.

According to the dual expression thesis, when a subject self-ascribes a mental state to herself by avowing it from the first-person perspective (e.g., "I believe that p", "My tooth hurts", etc.), the avowal expresses both a first-order mental state (e.g., my belief that p or the pain in my tooth) and a second-order belief that the subject has *in a qualified sense*. This qualified sense cannot be the one characteristic of the opining sense because first-person avowals are made on the basis of no specific evidence about what is considered to be their subject matter (i.e., the subject's mental states), and it is characteristic of beliefs in the opining sense that they are formed on the basis of specific evidence about their subject matter. Thus, according to Bar-on, first-person avowals express second-order beliefs that the subject has in the holding-true sense because if a subject who has the disposition to avow a mental state from the first-person perspective were asked whether she has that mental state, she would immediately assent on the basis of no specific evidence about her current mental states. For instance, if a subject has the disposition to avow "I feel thirsty" from the first-person perspective and she is asked whether she is thirsty, she will immediately answer affirmatively ("Yes, I feel thirsty") on the basis of no specific evidence about whether she is thirsty or not, and so, she has the second-order belief that she is thirsty in the holding-true sense.

Therefore, according to the dual expression thesis, when subjects issue a first-person avowal, they express both the self-ascribed mental state and the relevant second-order belief (which they have in the holding-true sense). For instance, the first-person avowal "I feel thirsty" express both my thirst (i.e., first-order mental state) and my second-order belief that I am thirsty (a second-order belief that I have in the holding-true sense). As a result, according to Bar-on, subjects who issue a first-person avowal are in the appropriate epistemic position to have a belief about whether they have the mental state self-ascribed in the first-person avowal (i.e., in the appropriate epistemic position to have a second-order belief).

Secondly, first-person avowals can be *true or false*, and so, the corresponding second-order beliefs expressed by them can be true or false as well. When the subject has the mental state self-ascribed in the first-person avowal, the first-person avowal is true, and so it is the corresponding second-order belief expressed by the first-person avowal. For instance, my first-person avowal "I feel thirsty" and my corresponding second-order belief expressed by that first-person avowal are true if I am thirsty as a matter of fact. By contrast, when the subject doesn't have the mental state self-ascribed in the first-person avowal, the first-person avowal is false, and so it is the corresponding second-order belief expressed by the first-person avowal. Bar-on gives the following example of false first-person avowal (Bar-on, 2004, p. 322).

Imagine that I'm in the dentist's chair and I'm about to have one of my teeth fixed. So, I open my mouth, I see the dentist introducing the drill into my mouth, and I shout "I feel a terrible pain in my tooth!" (or: "My tooth hurts so much!"). However, it happens that the dentist didn't reach my tooth yet. According to Bar-on, the first-person avowal "I feel a terrible pain in my tooth!" is a false self-ascription of pain in this case because, even if it expresses *the* mental state of pain, it doesn't express *my* pain insofar as I don't actually have any pain (for the dentist didn't reach my tooth). In Bar-on's terms, the first-person avowal "I feel a terrible pain in my tooth!" is false in the Exp$_3$ sense (i.e., in the semantic sense) because the first-person avowal doesn't Exp$_2$ (i.e., express in the causal sense) the self-ascribed mental state of pain (for I don't have any pain), even though it Exp$_1$ (i.e., express in the action sense) *a* mental state of pain only that *not my pain* (for, even if I don't have pain, the avowal semantically expresses pain because of the kind of expressive tool or vehicle that it is: a self-ascription of pain).

Finally, it remains to be seen how Bar-on's account explains the warrant of the second-order beliefs that are supposed to be involved in first-person self-knowledge. Bar-on's argument seems rather complex and confusing at this point. On the one hand, she seems to consider that the first-order mental state that is expressed and self-ascribed in a first-person avowal is both the *rational reason* that entitles the subject to make the avowal and the *epistemic reason* that warrants the second-order belief expressed in the avowal[21]. For instance, take the first-person avowal "I have a terrible toothache!". If I actually have a toothache (and I must have it if I am due to have first-person self-knowledge), the pain itself (i.e., the first-order mental state) is both the rational reason that entitles me to avow "I have a terrible toothache!" from the first-person perspective and the epistemic reason that warrants my second-order belief that I have a toothache. On the other hand, Bar-on seems to think that, because of the latter fact about first-person avowals, first-person avowals are more likely true than other (self)-ascriptions of mental states (e.g., third-person avowals or ascriptions of mental states made by others about me) because only first-person avowals express the same mental state that they

[21] "On the present proposal, what is epistemically unique about avowals is that the very same thing—one's being in M—provides both a rational reason for the avowal understood as an (expressive) act and an epistemic reason for the avowal understood as representative of the subject's self-judgment. An avowal as product (i.e., the self-ascription, or the judgment it expresses$_3$) requires no other warranting reason than whatever gives reason for the avowal as act (on the given occasion), thereby rendering it an act of expressing$_1$ the subject's M. It will thus turn out that whatever grounds avowals as expressive acts is also what allows them to represent a genuine and unique kind of knowledge. Although one can obtain genuine knowledge of others' states of mind, and can be justified in various ways in having beliefs about one's own states of mind, *only when avowing can one's epistemic warrant be the same as the rational cause of one's behavior*." (Bar-on, 2004, pp. 390-391).

self-ascribe[22]. As a result, the second-order beliefs expressed in first-person avowals are supposed to be strongly warranted because they are the result of a highly reliable expressive procedure of self-ascriptions of mental states: the expressive procedure of first-person avowals, which express both a first-order mental state and the corresponding second-order belief (dual expression thesis).

Therefore, according to Bar-on's neo-expressivist account, subjects who issue a first-person avowal on the basis of no evidence about what is considered to be its subject matter (i.e., the subject's mental states) might be in the appropriate epistemic position to have first-person self-knowledge. For they are supposed to be in the appropriate epistemic position to have a *true strongly warranted second-order belief*. Indeed, first-person avowals (e.g., "I believe that it is raining") are self-ascriptions of mental states that express both the first-order mental state that they self-ascribe (e.g., the belief that it is raining) and the second-order belief that one has such a first-order mental state (e.g., the belief that I believe that it is raining). Since that second-order belief is the result of a highly reliable expressive process (i.e., an expressive process that normally gives rise to true beliefs) because first-person avowals (normally) express the truth-maker (i.e., the first-order mental state) of the second-order belief, such a second-order belief enjoys a strong type of warrant that is exclusive of first-person self-knowledge. As a result, it is concluded that first-person self-knowledge is a "genuine and unique kind of knowledge" (Bar-on, 2004, p. 390) different from knowledge of the world (i.e., having a true warranted belief about a fact of the world), different from knowledge of other people's mental states (i.e., having a true warranted belief about someone else's mental state), and different from third-person knowledge of one's own mental states (i.e., having a true warranted second-order belief about one's own mental states by self-inspection). For the warrant of these latter beliefs is not of that stronger type insofar as they are not the result of such a highly reliable expressive process of belief-formation.

Some (Chrisman, 2009; Gertler, 2011) have questioned the capability of neo-expressivism to coherently account for the alleged *epistemic features* of first-person self-knowledge. However, the main mistake of neo-expressivist accounts in regard to self-knowledge is to take first-person self-knowledge as an epistemic phenomenon (*knowing that*) in need of explanation instead of as an expressive phenomenon (*knowing how*) in need of

[22] "So, […], true mental self-ascriptions produced in avowing do not merely happen to be true. There is a highly regular (though not exceptionless) correlation between avowing a mental state and being in the avowed mental state; thus, there is a highly reliable correlation between engaging in acts of avowing and producing true mental self-ascriptions." (Bar-on, 2004, p. 389).

explanation. For as it will be argued in the next section, the idea itself of first-person *epistemic* self-knowledge is conceptually flawed.

2.2.2 The impossibility of first-person epistemic self-knowledge

The idea that we can have first-person epistemic self-knowledge (*knowing that*) of our mental states is the idea that 1) there is a certain mental item of mine (i.e., either a mental state or an *aspect* of a mental state) that 2), in normal circumstances[23], can be accessed exclusively by me because 3) I have a way to access it that is exclusively mine (e.g., through the beliefs delivered by first-person processes such as: *bypass* —Fernández, 2013—, *doxastic schema* —Byrne, 2018; Gallois, 1996—, *deliberation* —Boyle, 2011; Moran, 2001—, or *first-person avowals* —Bar-on, 2004; Finkelstein, 2003—). Then, the idea of first-person epistemic self-knowledge involves the idea of *exclusive first-person access* to a mental item of mine. Also, since we can have access to other people's mental states by perceiving their expressive behaviour in appropriate conditions, the idea that we have *exclusive first-person access* to one's own mental states (or to some of their aspects) entails a relational view of expression. For if I had *exclusive first-person access* to a mental item of mine, mental states would be something more than expressive behaviour (which can be accessed by other people's perception when the conditions are appropriate and not exclusively by me), and so, *that* something more would have to be a further item to which a set of expressions would be related in some way (i.e., depending on how it is understood that expressions are evidence of mental states).

The following argument against the possibility of exclusive first-person access has been made (García, 2019b). This argument has been called *the argument from the replacement of self-reports* and it could be glossed out as follows:

1) It is a condition of access (in general) that the object accessed by the subject must be *ontologically robust*; i.e., available to access on different occasions. Perceptual access is an example of this. In paradigmatic cases of perception, it is possible to

[23] Not using brain scanners, for example.

have perceptual access to a single object on different occasions and by different subjects. For instance, my sister and I can have perceptual access to the book that I have on my desk on multiple occasions as long as the book is still on my desk. We can have access to the book now before leaving the house and later after coming back, now before closing our eyes and later after opening them up, or now before turning the lights off and later after turning them on again. So, the book on my desk is ontologically robust.

2) The mental items (i.e., mental states or aspects of mental states) to which I am supposed to have *exclusive first-person access* cannot be ontologically robust; i.e., available to access on different occasions. Cases of exclusive first-person access are considered to be reported by *self-ascriptions*, such as "I want a piece of cake" or "I have a pain in my leg". Those self-ascriptions are supposed to aim at *accuracy*. For when a subject changes a self-ascription for another, the change is seen either as the effect of a change of mind that allegedly makes the current self-ascription inaccurate (e.g., "I want a piece of cake. Wait a moment, I changed my mind, I would prefer an ice cream") or as the effect of having found out that the current self-ascription was inaccurate all along (e.g., "I have a pain in my leg. Wait a moment, it is an itch more than a pain"). However, since the subject is supposed to have *exclusive* first-person access to the object of the self-ascription (i.e., the mental item), there is no *standard of accuracy* independent of the self-ascription itself. A self-ascription will be accurate *by default* in each case only until the subject replaces it for a new self-ascription (hence the name of the argument: *from the replacement of self-reports*)[24]. However, if no independent standard of accuracy can be found, the mental item to which the subject allegedly has exclusive first-person access cannot be ontologically robust. Imagine that a subject keeps issuing the same self-ascription (e.g., "I want ice cream") on multiple occasions. Insofar as self-ascriptions are accurate in each case *by default* only until they are replaced by a different one, there is no reason to think that the subject has exclusive first-person

[24] It could be objected that there is an independent standard of accuracy: that the expressive behaviour of the subject *matches* the private mental item that she self-ascribes to herself. However, insofar as the self-ascription is of a *private* mental item (i.e., an item that can exclusively be accessed by the subject), the subject's expressive behaviour (which is available to other people's perception) can't ever be the standard of accuracy of the self-ascription. The private mental item and the public expressive behaviour are considered to be different items (so that they can take place separately) and the self-ascription is supposed to be (also) about the private mental item and not (only) about the expressive behaviour.

access to the *same item* (e.g., the desire to eat ice cream) on those different occasions (e.g., each time that she says "I want ice cream") because there is no reason to think that he has access to *any item* at all: her self-ascription will be accurate by default until she changes it for another one, and she can change it for another one at any time because there is no standard of accuracy other than the self-ascription itself.

3) Hence, the idea of exclusive first-person access is conceptually flawed. The idea of access requires access to an ontologically robust item, but the idea of *exclusive* first-person access is incompatible with the idea of access to an ontologically robust item because it is incompatible with any independent standard of accuracy.

Therefore, first-person epistemic self-knowledge cannot exist. The idea of first-person epistemic self-knowledge involves the idea of exclusive first-person access. Since the idea of exclusive first-person access is conceptually flawed, the idea of first-person epistemic self-knowledge is conceptually flawed as well. As a result, the asymmetry between first-person avowals issued from the first-person deliberative perspective and third-person avowals issued from the third-person self-inspective perspective cannot be understood *epistemically*, namely, as the difference between self-ascriptions of mental states that enjoy a stronger type of warrant because they are the result of a special type of exclusive first-person process (e.g., bypass, doxastic schema, deliberation or first-person avowals) and self-ascriptions of mental states that doesn't enjoy that stronger type of warrant because they are not the result of an exclusive first-person process. By contrast, the difference between first-person avowals and third-person avowals should be understood *semantically*: first-person avowals are not self-ascriptions of mental states while third-person avowals are.

As the behavioural-expressivist account of Transparency proposes, the difference between first-person and third-person avowals is that first-person avowals are *expressive episodes* of first-order mental states (rather than self-ascriptions) while third-person avowals are *self-ascriptions* of first-order mental states (rather than expressive episodes of first-order mental states). On the one hand, first-person avowals (e.g., "I believe that it is raining", "I want chocolate" or "I have a headache") are expressive episodes of the avowed first-order mental state (e.g., belief, desire or headache) whose intentional content is about the world (e.g., the rain, the chocolate or my head) and not about my own mental states. So, if first-person avowals are made on the basis of any evidence at all (which is the case with first-person avowals of attitude), the evidence is about the fact of the world that they are about and not about one's own mental states. By contrast, third-person avowals (e.g., "I believe that it is raining", "I want chocolate" or "I have a pain in my leg") are *self-ascriptions* of first-order mental states (e.g., belief, desire or pain). So, since third-person avowals are *self-ascriptions* of first-order mental states, third-person avowals are *expressive episodes* of the avowed *second-order belief*, whose intentional content is about one's own mental states. That's why third-person avowals are made on the basis of evidence about one's own mental states (i.e., how one acts, how one feels or how one judges). Therefore, avowals cannot be *first-person self-ascriptions* of mental states: when they are first-person avowals, they are not self-ascriptions; and when they are self-ascriptions, they are not first-person avowals.

So, how does the behavioural-expressivist understanding of the asymmetry between first-person and third-person avowals compare with the neo-expressivist understanding that we saw in the last section? Both neo-expressivist accounts and the behavioural-expressivist account consider that first-person avowals are *groundless* and *authoritative* in regard to the mental state explicitly mentioned in the utterance, although they understand these properties of avowals in different ways: neo-expressivist accounts understand them in an epistemic way while behavioural-expressivist understand them in an expressivist way. It was explained how neo-expressivist accounts understand the groundless and authoritative character of avowals, so let's see how the behavioural-expressivist account explains them. On the one hand, the behavioural-expressivist account considers that first-person avowals are groundless *only* in the sense that they are issued on the basis of no specific evidence about the subject's mental states and *not* in the sense that they are issued on the basis of no specific evidence about their subject matter (for their subject matter is considered here to be an aspect of the world rather than the subject's mental states). The explanation is that avowals are expressive episodes of mental states (just as a smile of happiness or a cry of pain) and not self-ascriptions of mental states. So, if any evidence is involved in a first-person avowal (which in avowals of attitude is involved), it will be evidence about the aspect of the world that the intentional content of the avowal is about (i.e., its subject matter) and not about the mental states of the avowing subject. On the other hand, the behavioural-expressivist account considers that first-person avowals are authoritative *only* in the sense that if you want to know my mental states, I am the best person to ask, and *not* in the sense that first-person avowals enjoy a special kind of warrant that makes them more likely true. If you want to know my mental states, it is normally better to ask me than to ask a friend of mine for the same reason than, if you want to know whether it is raining, it is normally better to open the window and look than to check the weather on your smartphone's app. My answer is criterial evidence of my mental state (insofar as it is an expressive episode of my mental state) instead of symptomatic evidence (as the answer of my friend is); likewise, some drops of water falling on the street are criterial evidence of rain (insofar as they are aspects of the event of rain) instead of symptomatic evidence (as the weather app is).

In the two following sections, the behavioural-expressivist accounts of first-person and third-person self-knowledge are going to be explicated. Firstly, it is going to be argued that first-person self-knowledge is an expressive phenomenon that has to do with the way in which subjects express their mental states (*knowing how*). Afterwards, it is going to be argued that third-person self-knowledge is an epistemic phenomenon that has to with having a true warranted second-order belief (*knowing that*), and that subjects can sometimes acquire genuine *strongly warranted* second-order beliefs from the third-person perspective of self-inspection.

2.2.3 First-person expressive self-knowledge

Subjects express their mental states from the first-person perspective by exercising activities that express those mental states (e.g., talking, picking up the umbrella, crying out of pain), and so, first-person self-knowledge is a matter of how subjects exercise the different activities that express their own mental states (*knowing how*) and not a matter of how subjects form true warranted beliefs about their own mental states (*knowing that*). Since there are two senses of *knowing how*, there are two senses in which subjects can have or lack first-person self-knowledge depending on how they express their mental states from the first-person perspective. On the one hand, 1) a subject has or lacks first-person self-knowledge of a *mental state* depending on whether she has the *ability* to express that mental state. On the other hand, 2) a subject has or lacks first-person self-knowledge of an *expressive episode* of mental state depending on whether she exercises (on the given occasion) the ability to express that mental state in a *self-conscious* or in an *un-self-conscious* way. Let's explain these two senses of first-person self-knowledge in order.

Firstly, 1) subjects have first-person self-knowledge of a mental state in the *knowing how* sense of having the ability to express that mental state when they are able to express that mental state in an *appropriate way* in normal circumstances (i.e., being awake, not being too tired, not being drunk, etc.). Of course, making some mistakes (expressive failures) at expressing a mental state in normal conditions is compatible with having the ability to appropriately express that mental state, but the kind and number of those mistakes should be analogous to the kind and number of mistakes that subjects who master that ability could make. (Likewise, having the ability to read is compatible with misreading some words every now and then, but the kind and number of mistakes should be analogous to the kind and number of mistakes that subjects who master the ability to read could make). By contrast, subjects don't have first-person self-knowledge of a mental state in the *knowing how* sense of having an ability when they are not able to express that mental state in an appropriate way in normal circumstances (i.e., being awake, not being too tired, not being drunk, etc.) so that they don't qualify as having the ability to express that mental state. Let's see two examples of lack of first-person self-knowledge in the sense of not having the ability to express a mental state (which involves expressing it in an inappropriate way most of the time and in normal conditions).

The first example is about a subject who uses the wrong *episode of the expression* and the second example is about a subject who uses the wrong *vehicle of expression*. Imagine a subject who is so grateful to her neighbours for calling the police when they saw someone trying to break into her house that, after three months, she still explicitly thanks them whenever she runs into them, which happens numerous times (e.g., "Thanks a lot for calling the police", "I am very grateful to you", "Thanks to you nobody broke into my house", and so on). Given the community of the subject and the social contexts in which she expresses her gratitude, those are inappropriate expressive episodes of gratitude, as it is shown by the fact that the neighbours themselves find odd and inappropriate that she thanks them each time that they see each other. Given her community and social context, after thanking them a few times, she should have instead expressed her gratitude in different ways (e.g., being more attentive with them, engaging in small-talk with them, bringing them some homemade food, and so on). As a result, she doesn't have the ability to express her gratitude because she doesn't know how to express it in an appropriate way (given her community and social context); i.e., using the appropriate *episodes of expression*. Imagine now a subject who is nostalgic for her years as a university student, and so, she displays the expressive pattern of nostalgia when she remembers those years (e.g., she smiles in a certain way, she idealizes that period of time telling of only the good things, she says “That was the happiest period of my life”, etc). However, instead of avowing her nostalgia by saying “I am *nostalgic* for my years as a university student”, which is an appropriate expressive vehicle of nostalgia, she avows her nostalgia by saying “I am *depressed* because my years as a university student are gone” because, due to her lack of conceptual dexterity regarding the concept of nostalgia, the wrong expressive vehicle “I am depressed because of […]” constitutes an expressive episode of her defective instantiation of the pattern of nostalgia. As a result, the subject doesn't have the ability express her nostalgia because she doesn’t know how to express it in an appropriate way (given her community and social context); i.e., using the appropriate *vehicle of expression*.

Of course, both the subject who expresses her gratitude with the wrong episode of expression (given her community and social contexts) and the subject who express her nostalgia with the wrong vehicle of expression (given her linguistic community) can improve their ability to express their mental state and end up mastering that ability to the degree that is considered to be sufficient to qualify as having such ability. In the first case, the subject might notice that something is wrong in the reactions of her neighbours when she keeps explicitly thanking them. In the second case, the subject might read a novel with a nostalgic character or a friend could tell her that she is nostalgic rather than depressed. Anyway, if they manage to end up mastering their ability to express their mental state to the degree that is required to qualify as having that ability, they would acquire first-person self-knowledge of their mental states in the *knowing ho*w sense.

Secondly, 2) subjects have or lack first-person self-knowledge of the *expressive episodes* of a mental state in the *knowing how* sense depending on whether the expressive episodes in question are instances of *self-conscious expression* or instances of *un-self-conscious expression*. Indeed, the different steps involved in the exercise of the activity that expresses a mental state can be performed self-consciously or un-self-consciously, and so, the different episodes of expression involved in the exercise of the activity that expresses a mental state can be episodes of self-conscious expression or episodes of un-self-conscious expression. When a subject expresses a mental state self-consciously, she expresses that mental state *knowing what she is up to* during the exercise of the activity that expresses that mental state. For instance, a person giving a talk about the causes of economic crises is self-consciously expressing her beliefs about the causes of economic crises because she knows what she is up to during the performance of the different steps of the activity that expresses her beliefs (i.e., the talk). She speaks paying attention to the sentences that she pronounces, to the ideas that she has to explain first and later, to the reactions of the audience, she knows what she just finished explaining and what she has to explain now, and so on. As a result, she expresses her beliefs about the causes of economic crises with self-conscious episodes of expression. Notice that from the fact that the subject's beliefs about the causes of economic crises are self-consciously expressed on the given occasion (i.e., while she is giving the talk), it doesn't follow that they are always self-consciously expressed, for the subject might un-self-consciously express her beliefs on other occasions (e.g., she might distractedly nod in assent to what another person is saying in a different talk also about the causes of economic crises).

By contrast, when a subject expresses a mental state un-self-consciously, she expresses that mental state *without knowing what she is up to* when she performs the activity that expresses the mental state. For instance, a person can un-self-consciously express her intention to pick up the umbrella before leaving the house by picking up the umbrella distractedly and without paying attention before leaving the house, only to find out later that she indeed picked up the umbrella (fortunately) and that she left it on the co-pilot seat. As a result, she expresses her intention to pick up the umbrella with an un-self-conscious episode of expression (for when she picked up the umbrella without paying attention, she didn't know what she was up to). Notice that from the fact that the subject's intention to pick up the umbrella is un-self-consciously expressed (i.e., picking up the umbrella distractedly) on the given occasion, it doesn't follow that such intention is always un-self-consciously expressed, for the subject might have self-consciously expressed her intention on other occasions (e.g., she might have self-consciously expressed her intention to pick up the umbrella before by attentively leaving the umbrella in a visible place close to the door an hour before leaving the house).

Therefore, a subject has first-person self-knowledge of an expressive episode of mental state in the self-consciousness sense (*knowing how*) when that expressive episode is self-consciously expressed (i.e., knowing what one is up to); and a subject lacks first-person self-knowledge of an expressive episode of mental state in the self-consciousness sense (*knowing how*) when that expressive episode is un-self-consciously expressed (i.e., without knowing what one is up to).

Thus, according to the behavioural-expressivist account, first-person self-knowledge is a matter of exercising the activity that expresses a mental state in a certain way (*knowing how*) rather than a matter of having a true warranted belief (*knowing that*). Since there are two senses of *knowing how*, there are also two senses in which a subject can have first-person expressive self-knowledge. On the one hand, subjects have first-person self-knowledge of a *mental state* when they have the ability to express that mental state, that is, when they are able to appropriately express that mental state in adequate circumstances most of the times. On the other hand, subjects have first-person self-knowledge of an *expressive episode* when that expressive episode is self-consciously expressed (i.e., knowing what one is doing) in the context of the activity that expresses that mental state. Notice that lack of first-person self-knowledge of a mental state in the sense of not having the ability to express that mental state involves a lack of first-person self-knowledge of the episodes of expression of that mental state in the sense that they cannot be self-consciously expressed (i.e., knowing what one is up to). For one cannot express a mental state self-consciously (i.e., knowing what one is up to) if one doesn't have the ability to express that mental state in an appropriate way. Furthermore, since to have or to lack the ability to express a mental state is a matter of degree (i.e., one can perform an activity better or worse independently of whether one is able to perform that activity well enough to qualify as having the ability), to be able to express a mental state self-consciously is also a matter of degree (one is able to express a mental state more self-consciously or less self-consciously depending on the degree with which one masters the ability to express that mental state).

In the next section, it is going to be argued that third-person self-knowledge is an epistemic phenomenon (*knowing that*) and that self-conscious expression plays an important role in third-person self-knowledge; in fact, so important that it is related to the possibility of having a strongly warranted belief about one's own mental states (in the sense of having a higher *degree* of warrant and not in the sense of having a stronger *type* of warrant) from the third-person perspective of self-inspection.

2.2.4 Third-person epistemic self-knowledge

So far it has been argued that first-person self-knowledge is an expressive phenomenon. Thus, it is now necessary to explain what is third-person epistemic self-knowledge. According to the behavioural-expressivist account, a subject has third-person self-knowledge when she has a true second-order belief warranted on the basis of evidence about her own mental states. This evidence can be of two types: criterial or direct evidence (e.g., her feelings, thoughts, imaginings, behaviour, gestures, facial expressions, etc.) and symptomatic or indirect evidence (e.g., what her friends say about her, etc.). The second-order beliefs involved in third-person self-knowledge are formed by the third-person process of *self-inspection* and they are expressed in third-person avowals (e.g., "I believe that it is raining", "I am thirsty", etc.). For third-person avowals are expressive episodes of second-order beliefs that consist in judgements about the mental states of the avowing subject (i.e., they are self-ascriptions). The reason why self-inspection is considered to be a third-person process to know one's own mental states is that it consists in applying to one's own case the same *epistemic method* that we normally use to know other people's mental states: to deliberate and to make a judgement about someone's mental states on the basis of evidence about the subject's mental states (although in self-inspection that someone happens to be me).

However, even if self-inspection is a third-person method, *sometimes* it can deliver true (second-order) beliefs about my own mental states whose warrant is stronger in degree than the beliefs of other people about my mental states can possibly be *in that particular situation*. Indeed, when I self-inspect myself to make a judgement about my mental states, I can use *memory* and *introspection* as epistemic sources of evidence about my mental states in a way that they are independent of *perception*, whereas other people can only use their epistemic sources of evidence about my mental states in a way that they are dependent of *perception*. Therefore, it will be argued in this section, when I self-inspect myself to make judgements about my mental states, memory and introspection sometimes can present my expressive episodes in a *clearer way* to myself than perception can present my expressive episodes to others in that particular situation. As a result, it will be argued, in the absence of any *setback* that could specifically affect self-inspection by affecting the gathering and assessment of the evidence about my own mental states (e.g., a failure of memory, a failure of introspection, a motivated bias, etc.), it is sometimes possible for me to acquire by self-inspection true (second-order) beliefs about my mental states whose warrant is stronger in degree than the warrant of other people's beliefs about my mental states can possibly be in that particular situation. In other words, it is sometimes possible for me to acquire *third-person authoritative self-knowledge* (i.e., true second-order belief warranted to a higher degree than the warrant of other people's beliefs about my mental states can possibly be in that particular situation).

Notice, however, that the warrant of my self-inspective (second-order) beliefs about my own mental states is always of the same *type* as the warrant of other people's beliefs about my mental states. For self-inspection consists in applying the same epistemic method to my own case that people normally use to make judgements about third-person subjects (i.e., to deliberate and to make up one's mind about someone's mental states) and on the basis of the same type of evidence (i.e., someone's episodes of expression). Therefore, the warrant of my self-inspective second-order beliefs can only be stronger in *degree* (not in type) and only in those particular cases in which memory and introspection present my expressive episodes (i.e., evidence) in a clearer way than other people's perception can present my expressive episodes to them (at that moment).

In order to explain third-person epistemic self-knowledge by self-inspection, I am going to use the following two examples of *self-interpretation*; i.e., cases in which a subject says something that interprets the expressive content of her own behaviour:

> Imagine that I am in the living room watching TV and I go to the kitchen to take an ice cream from the fridge. So I open the fridge, I rummage around the shelves, I open the different drawers to check them out, and I finally close the door of the fridge with a facial expression of annoyance while thinking "There isn't any ice cream left!" before going back to the sofa. Imagine that a day has passed and my sister, who was in the kitchen when I went there, asks me "Yesterday you were looking for something in the fridge, what were you looking for?". To which I answer: "I was looking for an ice cream".

> And,

> Imagine that I am talking with a friend about the different kinds of ways in which people can face and overcome the kind of breakup in which your partner leaves you for another person. Imagine that he says about me "You have a cold character in regard to affective relationships, so you wouldn't feel as bad as other people, would you?". However, I disagree with him, so I answer his question saying "I think I would feel terrible in that situation, even worse than if there weren't any third-person involved".

In these cases, my self-interpretation of the expressive content of my own behaviour in the past (e.g., opening the fringe, rummaging through the shelves, etc.) and my self-interpretation of the expressive content of my own behaviour in a hypothetical or counterfactual situation (e.g., feeling terrible in the situation of such a breakup) *might* have more weight or authority (i.e., they might not be easily questioned by others) than the third-person interpretations made by my sister or by my friend (which might be easily questioned by me). So, it is relevant to ask: how can I know (*sometimes*) the expressive content of my own behaviour in the past and the expressive content of my own behaviour in a hypothetical or counterfactual situation better than others?

Finkelstein (2003) and Cassam (2014) offer two different accounts of the phenomenon of self-interpretation. Finkelstein considers that self-interpretations are constitutive of a more "authoritative" kind of knowledge than third-person interpretations because they are *first-person* self-interpretations (i.e., self-interpretations made from the first-person perspective). By contrast, Cassam considers that self-interpretations are always *third-person* self-interpretations because they are instances of third-person *inferential* self-knowledge, and that they might have more weight than interpretations made by third-person subjects about me because they can be warranted on the basis, not only of what Cassam calls *external evidence* (e.g., my behaviour, what other people say about me —"You look tired"—, gestures, facial expressions, etc.), but also on the basis of what he calls *internal evidence* (e.g., judgements, mental promptings, feelings, emotions, mental images, etc.), which is a kind of evidence only available to me. Unlike Finkelstein and Cassam, I will argue that the behavioural-expressivist account can explain the examples described above claiming that they can be either cases of *first-person* self-interpretations (i.e., self-interpretations made from the first-person deliberative perspective) or cases of *third-person* self-interpretations (i.e., self-interpretations made from the third-person self-inspective perspective). I will use only the first example to discuss Finkelstein's and Cassam's views.

Finkelstein considers that self-interpretations have more weight or authority than third-person interpretations (i.e., interpretations about me made by others) because self-interpretations are *first-person* self-interpretations that elaborate or flesh out the expressive content of my behaviour in a way that third-person interpretations cannot do, and so, they are constitutive of an authoritative kind of knowledge. The reason why, according to Finkelstein, is that only first-person self-interpretations constitute a single "unit of intelligibility" (Finkelstein, 2003, p.105) with the episodes of expression that they self-interpret, meaning that they "make sense together, in the light of each other" (p. 109). This unit of intelligibility is supposed to be the characteristic feature of the first-person perspective, which is understood here as "a broader genus" (p. 111) than expression because it is a genus that is supposed to include both the subject's expressions and the subject's self-interpretations of those expressions. As a result, the first of the examples described above would be explained as follows. My self-interpretation (i.e., "I was looking for an ice cream") of the expressive content of my own behaviour when I opened the fridge, rummaged through the shelves and closed the fridge with a facial expression of annoyance is supposed to be more authoritative than any possible third-person interpretation that my sister could make about me because only my self-interpretation is a first-person self-interpretation that belongs to the same "unit of intelligibility" as my expressive behaviour itself.

However, the problem with Finkelstein's view is that considering that expressions and self-interpretations are two different species of the broader genus of the *first-person perspective* makes it difficult to see how the first-person perspective could be characterized. It is true that Finkelstein says that the broader genus characteristic of the first-person perspective is a "unit of intelligibility"; i.e., a unit of different species (i.e., expressions and self-interpretations) that "make sense together, in the light of each other". But these remarks don't seem to be clear enough to avoid the perplexities aroused by the question of how that unit of intelligibility could be understood in order to characterize the *first-person perspective*. By contrast, the behavioural-expressivist account can offer a clear characterization of the difference between the first-person and the third-person *perspectives*, as well as of the broader genus of the *first-person* (i.e., "me" rather than "you"). According to the behavioural-expressivist account, the difference between the first-person perspective and the third-person perspective is a matter of whether one expresses a first-order mental state (e.g., "I believe that it is raining", "Delicious!", "What a terrible headache!", making a grimace of pain, and so on) or whether one expresses a second-order belief formed by self-inspection (e.g., Tom's self-ascription "I don't actually believe that men and women are equal"), respectively. In turn, the genus of the *first-person* (i.e., "me" rather than "you") is identical to the species of *expression*: both when I express a first-order mental state (i.e., first-person perspective) and when I express a second-order belief formed by self-inspection (i.e., third-person perspective) I am the person who *expresses* his own mental states (i.e., either my first-order mental states or my second-order beliefs). As a result, according to the behavioural-expressivist account, the genus of the *first-person* can be appropriately characterized by the concept of expression; i.e., by characterizing the different ways in which a subject can *express* her mental states, including both expressing a mental state from the first-person perspective and expressing a second-order belief formed by the third-person self-inspective perspective.

An alternative to Finkelstein's view is to consider that self-interpretations are always third-person self-interpretations because they are made from the third-person perspective in which third-person epistemic self-knowledge can take place. Cassam considers that self-knowledge (both of our *attitudes* and of our *phenomenal states*) is always third-person epistemic self-knowledge (i.e., true warranted belief) because it is the result of third-person inferential self-interpretations made on the basis of (external and/or internal) evidence about one's own mental states. To illustrate this, Cassam describes the example of Katherine. Katherine suspects that she might want to have another child and she wants to know whether she really has the *desire* to have another child or not. So, she considers things such as her general taste for big families, her belief that it is not so difficult to raise a few happy and well-educated children, how many times she has imagined herself playing with a second child, how she feels when she helps her current toddler dress himself, how she feels when a friend tells her that she is having a baby, and so on. Thinking about all these things and taking into consideration all her *background knowledge* about herself and her *context*, she interprets (what, according to Cassam, involves inference) the feeling that she has when a friend tells her that she is having a baby as the feeling of envy, the feeling that she has when she helps her current child dress himself as the feeling of yearning for another baby, and so on. As a result, Katherine *infers* that she has the *desire* to have another child. If the conclusion is true and appropriately warranted, Katherine has third-person inferential self-knowledge of her desire to have another child.

Moreover, Cassam argues that the third-person inferential procedure characteristic of third-person epistemic self-knowledge can deliver beliefs about one's own mental states whose warrant is stronger in *degree* (but not in *type*) than the warrant of the beliefs of other people about one's own mental states can possibly be (Cassam, 2014, p. 150). For I can make inferences about my mental states on the basis, not only of the *external evidence* available to everybody when making judgements about my mental states (e.g., actions, facial expressions, utterances, etc.), but also on the basis of the *internal evidence* that is only available to me (e.g., thoughts, emotions, feelings, etc.) when I make judgements about my own mental states. Thus, not only Katherine has third-person inferential self-knowledge of her desire to have another baby but also her second-order belief can be warranted to a higher degree than the belief of her partner about Katherine's desire can possibly be (for Katherine's belief can be warranted not only on the basis of external evidence but also on the basis of internal evidence).

Therefore, Cassam could explain the first example described above as follows. My self-interpretation "I was looking for an ice cream" is an instance of third-person epistemic self-knowledge because it is a true inferential self-interpretation made on the basis of both internal and external evidence about what I did yesterday when I went to the kitchen. Indeed, I remember me opening the fridge, rummaging through the shelves and drawers, thinking "There isn't any ice cream left" and closing the fridge with a facial expression of annoyance. As a result, I can answer my sister's question "What were you looking for yesterday when you came up to the kitchen?" with the third-person inferential self-interpretation "I was looking for an ice cream" because I have inferentially formed a belief about the expressive content of my behaviour from yesterday (i.e., the desire to have an ice cream) on the basis of external and internal evidence about the occurrent episodes of mental states that I had when I went to the kitchen. Also, since my third-person inferential self-interpretation is warranted on the basis of both external evidence (e.g., my action of opening the fridge and rummaging around) and internal evidence (e.g., my thought "There isn't any ice cream left" or my feeling of annoyance), my third-person inferential self-interpretation is warranted to a higher degree than my sister's third-person interpretation of my expressive behaviour can possibly be (for my sister's interpretation can only be warranted on the basis of external evidence).

Cassam admits that there are cases of raw sensations that might not be self-known by inferential self-interpretation. For instance, a clear case of pain or a clear case of nausea. However, he doesn't think that such a possibility sets a problem for his inferential account of self-knowledge. On the one hand, Cassam says that self-knowledge of complex mental states (e.g., feelings, emotions, beliefs, desires, etc.) is not based on self-knowledge of simple sensations (e.g., pain or nausea), so that self-knowledge of simple sensations contributes little or nothing to self-knowledge of more complex mental states. On the other hand, Cassam says that even *some* cases of simple sensations can sometimes be self-known by inferential self-interpretation. For instance, sometimes is possible to doubt whether a particular sensation is a pain or an itch, and it is sometimes possible to interpret that sensation as being one thing or the other depending on one's contexts and background knowledge (e.g., whether I've been recently injured on that part of my body or whether I have a mosquito bite on there)[25].

However, Cassam's inferential account of self-knowledge has a problem. Not only is it unable to explain most cases of self-knowledge of simple or raw sensations, like pain or nausea, it is also unable to explain cases of self-knowledge of attitudes, as it is shown by Katherine's example itself. Let's see why. Katherine is supposed to have self-knowledge of her desire to have a child because she inferred so on the basis of different pieces of evidence, among which were the important facts that she *felt envy* when a friend told her about her pregnancy and that she *felt a yearning* for a baby when she was helping her child dress himself. In turn, she is supposed to have self-knowledge of that feeling of envy because she inferentially interpreted her *feeling* as a feeling of envy in the context of her conversation with a friend, and she is supposed to have self-knowledge of that feeling of yearning for a baby because she inferentially interpreted her *feeling* as a feeling of yearning for a baby in the context of helping her child dress himself. However, how did Katherine *know* that she had these feelings? It seems that Katherine's inferential self-interpretations from her own feelings require that Katherine has *non-inferential self-knowledge* of the fact that she had a feeling at all when a friend told her about her pregnancy and when she was helping her child dress himself. The question is important because those feelings (i.e., the feeling of envy and the feeling of yearning for a baby) are supposed to be among the contents from which she inferentially self-interprets herself to acquire third-person inferential self-knowledge of her desire to have another baby.

To avoid this problem, Cassam argues that it is not plausible to think that Katherine knows her feeling of envy and her feeling of yearning from their phenomenology, for there isn't any distinctive phenomenology of the different feelings that a subject could have:

[25] "If this is right [i.e., that there are cases of non-inferential self-knowledge of sensations like pain or nausea] then here we have a case of non-interpretative and non-inferential access to an 'internal prompting'. However, this is a possibility that inferentialist about self-knowledge of standing attitudes and more complex feelings and emotions can allow, as long as self-knowledge of simple sensations isn't seen as the basis of all other self-knowledge. Knowledge of sensations like pain contributes little to self-knowledge of standing attitudes, and even the true extent to which our access to so-called 'simple' sensations is non-interpretative can be questioned. The answer to the question 'Are you in pain?' isn't always obvious, and it's not unusual for people to report being conscious of sensations which they are unsure whether to classify as pain. In such cases, it can happen that discovering the cause of the sensation can help you to make sense of it, to classify it one way rather than another. Here, your access to the sensation look[s] genuinely interpretative." (Cassam, 2014, p. 164).

> "You can't just 'read off' from the way you feel that your yearning is for another child. You can yearn for any number of things, and it would be odd to think that each yearning has its own distinctive phenomenology. When you identify your feeling as the yearning for another child what you are doing is interpreting it, and your cognitive effort is the effort of interpretation. Crucially, when you interpret your feeling you don't just go on 'how it feels'. You also take account of contextual factors, such as the fact that you have recently been thinking about whether to have another child. More often than not, at least in the case of complex feelings and emotions, it is your knowledge of the context which makes it possible for you to determine its nature, which means that you are to some extent inferring what you feel from your background knowledge". (Cassam, 2014, p. 163).

Thus, Cassam seems to conclude that Katherine has inferential self-knowledge of her feeling of envy and of her feeling of yearning and that the content from which that inference is made cannot be other than the context and background knowledge of Katherine because feelings don't have a distinctive phenomenology. I agree with Cassam in which it is not plausible to think that one knows her feelings by identifying them from their phenomenology. However, that doesn't make Cassam's inferential account more palatable. Cassam distinguishes between external evidence (e.g., actions, facial expressions, utterances, what other people say about me, etc.) and internal evidence (e.g., feelings, passing thoughts, emotions, mental images, etc.) of our mental states (e.g., Cassam, 2014, p. 150). Since external evidence is acquired by *perception*, it is not problematic for Cassam's inferential account. For instance, I can see my facial expression in the mirror and infer that I am tired, I can infer that I like someone because I see that I spend a lot of time with her, and so on. The problem for Cassam's inferential account, though, arises in regard to internal evidence of mental states, like the *feelings* of Katherine. To make an inference is to go from one content to another content. One can infer C from B, but one cannot infer B from B. Then, Katherine can infer that she has the *desire* to have another baby from the fact that she has the *feeling of envy* when talking to her pregnant friend and the *feeling of yearning* when she helps her child dress himself. In this case, Katherine infers C (i.e., her desire to have another baby) from B (i.e., her feelings of envy and yearning). However, from which content is Katherine supposed to infer her feeling of yearning and her feeling of envy? Cassam has two possible answers for this question and both are problematic. Let's see them in order.

Firstly, (1) Cassam could argue that Katherine's feelings of envy and yearning are *inferential in themselves* because they *ontologically* depend on Katherine's background knowledge and context (i.e., the conversation with her pregnant friend and the act of helping her child dress himself) rather than depending on a distinctive phenomenology. However, to say so would be as much as to say that Katherine infers a content B (i.e., her feelings of envy and yearning) from the content B itself (i.e., her feelings of envy and yearning), which goes against the notion of inference. Secondly, (2) Cassam could say that Katherine infers her feelings of envy and yearning (i.e., the content B) only from her *background knowledge* and her *context* (i.e., the conversation with her pregnant friend and the act of helping her child dress himself), which are a content (i.e., the content A) different from the feelings of envy and yearning themselves (i.e., the content B). However, to say so is problematic for two reasons. On the one hand (2.1.), to say so would blur the difference, endorsed by Cassam, between internal and external evidence of mental states. All the evidence used by Katherine to infer her feelings of envy and yearning from her background knowledge and context would have to be external evidence about her own mental states (e.g., facial expressions, actions, bodily posture, demeanour, etc.) available to *perception*. For, since internal evidence is never available to perception (e.g., internal promptings, passing thoughts, etc.), if Katherine needed internal evidence to infer her feelings of envy and yearning, the question about the content from which Katherine is supposed to infer that internal evidence (e.g., internal promptings, passing thoughts, etc.) would be triggered again (and *at infinitum*). Thus, *all* the evidence from which Katherine infers her own mental states (e.g., feelings of envy and yearning) would have to be *external* evidence at the end of the inferential chain (e.g., facial expressions, actions, bodily posture, demeanour, etc.), and *perception* would have to be the ultimate source of evidence of Katherine's inferential self-knowledge. However, this idea seems highly implausible: Katherine doesn't need to *look* at her facial expression in a mirror or to *listen* to what she is saying while talking with her pregnant friend to infer on that basis (i.e., external evidence) that she has the feeling of envy. On the other hand (2.2.), nothing in Cassam's account seems to rule out the possibility of the fact that Katherine could have exactly the same conversation with her friend and exactly the same facial expression (i.e., external evidence) *with and without* the feeling of envy, or the possibility of the fact that she could help her child dress himself in exactly the same way (i.e., external evidence) *with and without* the feeling of yearning for another baby. Thus, nothing in Cassam's account is able to explain how Katherine's feelings of envy and yearning could be inferred by Katherine herself only from external evidence (e.g., her tone of voice, her actions, her facial expression, and so on).

As a result, Cassam's inferential account cannot explain how Katherine has self-knowledge of her feelings of envy and yearning, and so, Cassam's account is not able to explain Katherine's third-person epistemic self-knowledge of her desire to have another baby. Therefore, Cassam's inferential account not only is unable to explain how we can have self-knowledge of our raw feelings and sensations (as Cassam himself admits) but also how we have self-knowledge in cases that Cassam considers paradigmatic of third-person inferential self-knowledge (i.e., Katherine's case). Thus, to look for another account of third-person epistemic self-knowledge is in order.

Then, it is time to explain how I think that the behavioural-expressivist account should understand the process of self-inspection by which subjects are able to acquire third-person epistemic self-knowledge. Self-inspection can deliver third-person epistemic self-knowledge because it can deliver true second-order beliefs warranted on the basis of evidence about my mental states provided by *perception* (e.g., I might judge that I am tired by perceiving my tiredness in my facial expression —criterial or direct evidence— if it happens that I come across a mirror and I look at my face), *inference* (e.g., I might judge that I am thirsty by inference on the basis of the fact that I feel the urge to eat without being hungry —symptomatic or indirect evidence—, something that I found that sometimes happens to me when I am thirsty), *memory* (e.g., I might judge that I wanted an ice cream on the basis of the fact that I remember myself looking for an ice cream —criterial or direct evidence—) and *introspection* (e.g., I might judge that I would feel terrible in the hypothetical situation of a breakup on the basis of the fact that I feel some kind of despair when I imagine myself in that hypothetical situation with the aim of finding out how I'd feel —criterial or direct evidence—), all of which are epistemic sources of evidence that can provide me with evidence about my own mental states.

Moreover, as it will be shown below, *sometimes* self-inspection can deliver true beliefs about my mental states whose warrant is stronger in degree than the warrant that other people's beliefs about my mental states can possibly have *on the given situation*. For memory and introspection can be used as epistemic sources of evidence independent of perception only when self-inspecting oneself and not when making judgements about other people's mental states; and *sometimes* memory and introspection are able to present my expressive episodes in a *clearer way* to me than perception is able to present my expressive episodes to other people on the given occasion. As a result, even if self-inspection is a third-person process of self-knowledge because it consists in applying to one's own case the same *epistemic method* that we normally use to know other people's mental states (i.e., to deliberate and to make a judgement about someone's mental states) on the basis of the same type of evidence (i.e., someone's episodes of expression), it can sometimes deliver true second-order beliefs warranted to a higher degree thanks to memory and introspection (i.e., which can only be used as epistemic sources of evidence about my mental states independent of perception in my own case). Thus, self-inspection can sometimes deliver *third-person authoritative self-knowledge* (i.e., true second-order beliefs warranted to a higher degree than other people's beliefs about one's own mental states can possibly be on the given situation).

Before seeing how the behavioural-expressivist account can explain the two examples that were described at the beginning of this section, let's explain memory and introspection as expressive faculties first. On the one hand, *memory* is the expressive capacity to remember (some of) the self-conscious[26] episodes of mental state that one expressed in the past by currently reviving them in the form of episodes of *memory-states*. So, by exercising my memory, I can currently express episodes of memory-states whose intentional content is the self-conscious episodes of mental state that I (likely)[27] expressed in the past. For instance, a subject exercises her memory when she spontaneously says "[I remember] what a terrible pain I felt when I broke my leg!" to express her memory-state of the pain. Notice that the memory-state of the pain is *about* the pain, and so, it is something different from the mental state of pain that I self-consciously expressed in the past. Since the memory-state is a mental state different from the self-conscious expressive episode from the past that it is about, sometimes memory can fail: I can express a memory-state whose intentional content is about an episode of mental state that I didn't express in the past (at least, not in the way described by the memory-state). For instance, I can (falsely) remember that I performed the action of visiting the Alhambra when I went to Granada a few years ago (memory-state), when in fact I only saw the Alhambra in a travel guide in which I also read the description. On the other hand, introspection is the expressive capacity to imagine oneself in a counterfactual or hypothetical situation with the aim of expressing introspective-states whose subject matter or intentional content is about the episodes of mental state that I would (likely)[28] express in that hypothetical or counterfactual situation. For instance, I can exercise my introspection imagining myself in the hypothetical situation of a break-up and saying "[I introspect that] I'd feel terrible if she leaves me!" to

[26] Notice that it is not possible to remember un-self-conscious episodes of expression. For instance, if I distractedly pick up the umbrella without knowing what I was up to, I cannot later remember myself picking up the umbrella.
[27] For memory can fail.
[28] For introspection can fail.

express my introspective-state of despair (or *introspective-despair*). Notice that the introspective-despair is *about* the feeling of despair that I would (likely) express, and so, it is different from the genuine feeling of despair that I am supposed to express if the actual break-up occurred (as I might, unfortunately, find out if my partner actually leaves me). Since introspective-states are mental states different from the hypothetical or counterfactual episodes of mental state that they are about (for my genuine feelings about my hypothetical breakup don't currently exist), sometimes introspection can fail. Indeed, the hypothetical situation might end up taking place and I might end up having quite different mental states from the mental states pointed out by my introspective-states (e.g., my relationship might end as a matter of fact and I might find relief rather than despair).

Then, the behavioural-expressivist account could explain the examples described at the beginning of this section by claiming (against Finkelstein and against Cassam) that self-interpretations (e.g., "I was looking for an ice cream" or "I think I would feel terrible in that situation") can be issued either from the first-person deliberative perspective (i.e., first-person self-interpretations) or from the third-person self-inspective perspective (i.e., third-person self-interpretations). When self-interpretations are issued from the first-person deliberative perspective, they are either expressive episodes of *memory-states* whose intentional content is about the self-conscious episodes of mental state that I (likely) expressed in the past or expressive episodes of *introspective-states* whose intentional content is about the episodes of mental state that I would (likely) experience in a hypothetical or counterfactual situation. By contrast, when self-interpretations are issued from the third-person self-inspective perspective, they are expressive episodes of *second-order beliefs* formed by self-inspection on the basis of evidence about one's own mental states, and so, they might be instances of third-person epistemic self-knowledge (i.e., true warranted belief). Among the evidence on the basis of which I can make a third-person self-interpretation of the expressive content of my past behaviour are the memory-states provided by the expressive faculty of memory, and among the evidence on the basis of which I can make a third-person self-interpretation of the expressive content of my hypothetical behaviour are the introspective-states provided by the expressive faculty of introspection. Let's see how the behavioural-expressivist account can explain the two examples described at the beginning of this section in order.

The first example is explained by the expressive capacity of *memory*. My sister asks me the question "Yesterday you were looking for something in the fridge, what were you looking for?" and I answer with a self-interpretation of the expressive content of my behaviour in the past: “I was looking for an ice cream”. My self-interpretation seems to have more weight or authority (i.e., it seems to be presumably true) than any interpretation of the expressive behaviour than my sister could make about me in her current situation (e.g., “You were looking for an orange”, “You wanted to cook something but you didn’t find the ingredients” etc.). How can I know better than my sister what I was looking for yesterday? How can I know the expressive content of my behaviour from the past better than her?

On the one hand, I can answer my sister’s question from the first-person deliberative perspective by exercising my capacity of memory and expressing an episode of memory-state (whose intentional content is about my behaviour from the past) on the basis of no evidence about my *current* mental states. In this case, I answer with a first-person self-interpretation (i.e., an expressive episode of memory-state) about the expressive content of my behaviour from the past on the basis of no evidence about my current mental states. For instance, I can answer my sister’s question directly and on the basis of no evidence about my current mental states with the first-person self-interpretation “I was rummaging through the fridge looking for an ice cream” or “I got angry because there wasn’t any ice cream left!”. In these cases, my answers are *expressive episodes* of memory-states whose intentional content is the self-conscious episodes of the desire to have an ice cream that I expressed in the past when I went to the kitchen. As a result, in these cases, I might know better than my sister what I was doing yesterday when I went to the kitchen because I might exercise my expressive capacity of memory in an appropriate way (*knowing how*) and I might manage to express the *right* memory-state; i.e., the memory-state that describes the expressive content of my *actual* behaviour from yesterday. By contrast, my sister can exercise her memory only to remember herself *perceiving me* looking for something (i.e., the intentional content of her memory-states is always herself *perceiving* my expressive episodes), but she cannot exercise her memory to express memory-states whose intentional content is *my* self-conscious expressive episodes *themselves*. Also, since my sister only got to know that I was looking for something (without knowing what I was looking for) when she perceived me in the kitchen, her interpretations of my expressive behaviour cannot go further until I tell her that I was looking for an ice cream; for instance, with a first-person self-interpretation consisting in an expressive episode of memory-state. In this case, to know better than my sister is a matter of *first-person expressive self-knowledge*: exercising my expressive capacity of memory to express the right episode of memory-state, that is, the episode of memory-state whose intentional content is about *me expressing* the episodes of mental state that I actually expressed when I went to the kitchen.

On the other hand, I can answer my sister's question "Yesterday you were looking for something in the fridge, what were you looking for?" from the third-person self-inspective perspective as well; that is, making a judgement about the expressive content of my past behaviour on the basis of the evidence about my mental states from the past provided by my memory (i.e., on the basis of my memory-states). In this case, I answer with the third-person self-interpretation "I was looking for an ice cream" made from the third-person self-inspective perspective on the basis of the evidence provided by memory (i.e., memory-states) about the mental states that I had when I went to the kitchen. Indeed, imagine that, rather than answering my sister's question directly from the first-person deliberative perspective with an expression of memory-state, I stop for a moment to *think silently*[29] about what I did yesterday. So, firstly, I exercise my expressive capacity of memory to remember me opening the fridge, rummaging around, thinking "There isn't any ice cream left!" and closing the door. And then, after having silently exercised my expressive capacity of memory from the first-person deliberative perspective, I make a judgement from the third-person self-inspective perspective about the expressive content of my past behaviour on the basis of my memory-states. As a result, I can answer my sister's question with the judgement or third-person self-interpretation "I was looking for an ice cream". This judgement or third-person self-interpretation is an expressive episode of my *second-order belief* that the expressive content of my behaviour when I went to the kitchen was the desire to have an ice cream. Or, for short, it is an expressive episode of my *second-order belief* that yesterday I wanted to have an ice cream. However, how can I know the expressive content of my behaviour *better* than my sister? In this case, my *memory-states* present my expressive episodes to me as a self-inspective subject in a *clearer way* than my expressive episodes were presented to my sister's *perception*. For my memory-states present me as *self-consciously expressing* my episodes of desire (i.e., their intentional content is about *me* self-consciously performing the activity of looking for an ice cream in the fridge), while my sister *perceived* someone else's expressive episodes (and so, her memory-states presents *her perceiving* someone else's expressive episodes) in a perceptual context in which the precise expressive content of those episodes of mental state wasn't clear enough: she knew that I was

[29] Thoughts should be understood as forming part of the expressive content of some episodes of mental states and not as further items different from my expressive behaviour. For silent thoughts form part of the expressive content of the same *kind* of episodes of mental states as the episodes of mental states that are constituted by aloud utterances (which are a type of vehicle of expression). In the section 4.8. of this essay, it will be argued that thoughts are expressed in expressive episodes constituted by vehicles of expression such as gestures, bodily postures, facial expressions, a certain demeanour, and so on.

looking for something, but not what I was looking for ("Yesterday you were looking for something in the fridge, what were you looking for?").

As a result, my self-inspective second-order belief that yesterday I wanted an ice cream is warranted to a higher degree than my sister's belief about my expressive content could have been in her particular situation (i.e., before I tell her). Both my second-order belief and her belief about me are formed by a third-person epistemic method; i.e., making a judgement about someone's mental states on the basis of evidence about her mental states. And both my second-order belief and her belief about me are based on the same kind of evidence; i.e., on my expressive behaviour when I went to the kitchen. However, in this particular case, my *memory-states* presents my behaviour (i.e., the evidence) as clearly having the expressive content of my desire to have an ice cream, while my behaviour were presented to my sister's perception as having a *less clear* expressive content, namely, the expressive content of wanting *something*. Thus, the difference between the *warrant* that my self-inspective belief has in this case and the *warrant* that any belief of my sister could have had in the given circumstances (before I told her) is a matter of *degree* of strength and not a matter of *type*. Both my self-inspective belief and my sister's belief are formed by a third-person epistemic method (i.e., making a judgement on the basis of evidence) and they are warranted on the basis of the same type of evidence (i.e., my expressive episodes). However, my self-inspective belief enjoys a degree of strength that my sister's belief cannot enjoy in this case (before I tell her) because my memory presents the evidence (i.e., my expressive episodes) with a degree of clarity that my sister's perception couldn't reach in this case. Since the difference here between my sister and myself is a matter of *form of presentation* (i.e., with more or less clarity) of the same kind of evidence (i.e., my expressive behaviour)[30], and not a matter of *different kinds* of evidence (i.e., my expressive behaviour vs. a further item behind my expressive behaviour), no worries of defeating the non-relational project of the behavioural-expressivist account arise with this account of third-person epistemic self-knowledge.

Notice, however, that it is not always the case that memory presents one's own expressive episodes to oneself when performing self-inspection in a clearer way than to other

[30] Likewise, I can perceive the keyboard on which I type both by looking at it or by carefully touching the keyboard with my eyes closed. These perceptual experiences are two different forms of presentation of the same object, and so, if I make the judgment that there is a keyboard on my table on the basis of any of those perceptual experiences, I make a judgement on the basis of the very same type of evidence insofar as they are different *forms of presentation of the same object*. Alternatively, if I judge that there is a keyboard on my table on the basis of testimony, I make a judgement on the basis of a different type of evidence insofar as no presentation of the keyboard occurs (neither in one *form* nor in another).

people's perception. Imagine someone overtly and clearly expressing pain because she just broke her leg (i.e., shouting, crying, saying "It hurts a lot!", etc.). Imagine that a month has passed and the person who broke her leg is talking with a friend who saw how she broke her leg. In this case, even if the memory of the person who broke her leg presents her expressive episodes as clear episodes of pain because she remembers *herself self-consciously expressing her pain* (rather than herself *perceiving* behaviour of pain), the degree of clarity with which those expressive episodes are presented to the victim's memory is the same as the degree of clarity with which they were presented to her friend's perception: in both cases, the episodes are presented as clearly having the expressive content of terrible pain. As a result, both the perceptual belief of the friend and the self-inspective second-order belief of the person who broke her leg enjoy a similar degree of strength in their warrant because the expressive content of the victim's behaviour (i.e., the evidence) is presented with a similar degree of clarity to both subjects (i.e., the memory-states of the victim presents her own episodes of pain with the same degree of clarity that they were presented to her friend's perception).

The difference between this particular example of pain and the example of my sister has to do with the *conditions* in which expressive content of the subject's behaviour is perceived on the given occasion (i.e., it has to do with the *perceptual conditions*[31]). In this example of pain, the perceptual conditions in which the friend perceives the victim's behaviour of pain are optimal because he has perceptual access to enough *context* (e.g., how she felt down) and enough *portion* of the victim's expressive behaviour (e.g., how she shouts or cry while protecting her leg) to perceive its expressive content in a clear way: as clearly expressing a terrible pain. However, the perceptual conditions of my sister weren't as good as they are in this example of pain. My sister didn't perceive enough context (e.g., she couldn't see me picking up an ice cream insofar as I left the kitchen without finding one) and she didn't perceive enough portion of my expressive behaviour (e.g., neither she asked me at that moment "What are you looking for?" nor I told her "I am looking for an ice cream") to be able to have the appropriate perceptual conditions to have clear perceptual access to the expressive content of my behaviour: my desire to have an ice cream. As a result, while the expressive content of my behaviour was presented in an less clear way to my sister's perception (i.e., she remembers herself perceiving me looking for *something* in the fridge) than to my memory (i.e., I remember me clearly expressing my desire to have an ice cream), in the example of the pain the expressive

[31] More about the perceptual conditions of expressive content (which are different from the perceptual conditions of material items) in section 4.8.

content of the subject's behaviour is presented with the same degree of clarity to the victim's memory (i.e., she remembers herself clearly expressing her terrible pain) than to her friend (i.e., he remembers himself perceiving her expressing her terrible pain in a clear way).

However, is it really necessary to posit the possibility of third-person epistemic self-knowledge (i.e., true warranted belief) on the basis of memory? Cannot the example of my sister and me be explained exclusively as a case of first-person expressive self-knowledge? The fact that third-person epistemic self-knowledge based on evidence provided by memory is a genuine phenomenon can be shown by the following two examples. Firstly, I can sometimes remember something exercising my memory from the first-person deliberative perspective on the basis of no evidence about my current mental states, but refrain myself from judging that the intentional content of the memory-states that I express is true. For instance, if you ask me whether I've ever been in the Alhambra, I might remember me being there when exercising my memory from the first-person deliberative perspective, but refrain myself from judging that I was there because I question the truth of my memory-states: I always upload pictures to Facebook of the beautiful places that I visit when I travel and I don't remember having any picture of the Alhambra. As a result, I can answer your question about whether I've ever been in the Alhambra from the third-person self-inspective perspective (rather than from the first-person deliberative perspective) saying "I seem to remember that I was there, but I doubt it because I don't remember having any picture of the Alhambra on my Facebook account", or just saying "I don't know" if I don't want to give you the whole explanation. This judgement is made from the third-person self-inspective perspective only on the basis of the evidence about my expressive behaviour in the past that is provided by memory to me (i.e., only on the basis of memory-states whose intentional content is about me being in the Alhambra and about the fact that I don't have any picture on Facebook, respectively). I seem to remember that I was in the Alhambra when I went to Granada, but I also remember that I don't have any picture of the Alhambra on my Facebook account. So, I answer your question with the suspension of judgement "I don't know" made on the basis of evidence provided by memory (i.e., I answer with an expressive episode of my second-order belief that I have conflicting memories about what I did when I went to Granada).

Secondly, that I can form second-order beliefs about myself from the third-person perspective of self-inspection on the basis of evidence provided by memory is proved by the fact that second-order beliefs and memory-states have different patterns of expressive behaviour. For instance, thanks to the fact that I can form second-order beliefs about my mental states from the third-person self-inspective perspective on the basis of evidence provided by memory, the following situation can happen. I am watching TV and there is a report of how bad industrial ice cream is for health. I remember the numerous times that I went to the kitchen this summer looking for an ice cream (i.e., the numerous times that I had the intention to eat an ice cream) and I came back frustrated because there wasn't any ice cream left. So, I make the judgement "Thanks to my sister, who likes to shop only healthy grocery, I didn't eat more ice cream this summer than I should have". Clearly, this judgement is not *only* the result of exercising my expressive capacity of memory from the first-person deliberative perspective, for memory doesn't say anything about whether I tried (i.e., intention) to eat *more ice cream than I should*. By contrast, this judgement is the result of using both the evidence provided by memory and the evidence provided by the TV report. Thus, this judgement is an expressive episode of my second-order belief that this summer I tried (i.e., intention) to eat ice cream more times than I should have if I want to preserve my health. Therefore, the two examples described in the last two paragraphs show why third-person self-ascriptive judgements made on the basis of evidence provided by memory are a genuine phenomenon, and hence, why true second-order beliefs warranted on the basis of evidence about my mental states provided by memory (i.e., third-person epistemic self-knowledge) are a genuine phenomenon as well.

Now, it is time to explain the second of the examples described at the beginning of this section. The second example is explained by the expressive capacity of *introspection*. Imagine that I am talking with a friend about how a person might feel when your partner leaves you for another person after a long-term relationship. My friend asks me "Would you feel as bad as other people (with a character warmer than yours)?" and I answer this question with a self-interpretation of the expressive content of my behaviour in the relevant counterfactual or hypothetical situation; e.g., "I would feel terrible if my partner leaves me for another person". My self-interpretation can be a first-person interpretation resulting from exercising my expressive faculty of introspection from the first-person deliberative perspective or a third-person self-interpretation made from the third-person self-inspective perspective on the basis of the evidence about my hypothetical mental states provided by introspection. Then, how can I know better than my friend when I answer the question?

On the one hand, I can answer my friend's question on the basis of no evidence about my current mental states by exercising my expressive capacity of introspection from the first-person deliberative perspective; that is, with the expressive episode (e.g., "I would feel terrible if my partner leaves me for another person") of an introspective-state whose intentional content is about the expressive content of the behaviour that I would have in the relevant counterfactual situation. In this case, I might know better than my friend because I might have managed to exercise my expressive capacity of introspection in an appropriate way (*knowing how*) to express an episode of the *right* introspective-state; i.e., of the introspective-state whose intentional content is about the expressive behaviour that I would *actually* have if the relevant counterfactual situation occurred. By contrast, my friend cannot exercise his expressive capacity of introspection to express an introspective-state whose intentional content is about how *I* (but not him) would act in a certain counterfactual situation. The only thing that my friend can do is to make a judgement (*knowing that*) about how I would act in a counterfactual situation on the basis of his background knowledge about me (i.e., evidence about my mental states). Thus, in this case, to know better than my friend is a matter of *first-person expressive self-knowledge*: exercising the expressive capacity of introspection in an appropriate way to express the right episode of introspective-state; i.e., the introspective state whose intentional content is about the expressive behaviour that I would *actually* have if the counterfactual situation occurred.

On the other hand, I can answer my friend's question from the third-person self-inspective perspective on the basis of the evidence about my mental states provided by introspection (i.e., on the basis of my introspective-states) and on the basis of all my background knowledge about myself (which is provided by memory —memory-states—). In this case, I answer with the third-person self-interpretation or self-inspective judgement "I would feel terrible if my partner left me for another person", which is about the expressive content of my behaviour in a hypothetical situation and it is made on the basis of the evidence about my mental states provided by introspection and memory. Indeed, imagine that before answering my friend's question, I silently exercise my expressive capacity of introspection by imagining for a moment how I would feel if my partner left me for another person after a long-term relationship and my expressive capacity of memory to remember how I have actually felt in similar situations in the past (background knowledge). As a result, I answer with a self-inspective judgement "I would feel terrible if my partner left me for another person", which is a third-person self-interpretation about the expressive content of my behaviour in the relevant hypothetical situation and it is made on the basis of the evidence provided by introspection (e.g., that I would cry, that I wouldn't want to go out of my house, that I would feel despised, and so on) and my memory (e.g., how much time I have needed to recover from a breakup in the past or how I felt). This third-person self-interpretation or self-inspective judgement (e.g., "I would feel terrible if my partner left me for another person") is an expressive episode of my *second-order belief* that I would feel terrible if my partner left me for another person. However, how can I know better than my friend? I might know better than my friend in this case because my second-order belief (*knowing that*) might be warranted with a higher degree of strength than any belief than my friend could have about the issue in this particular situation (i.e., before I tell him the results of my introspective exercise). Indeed, whereas my second-order belief is formed on the basis of the evidence about my mental states provided both by introspection (e.g., that I would cry) and my memory (i.e., my background knowledge about me), the belief of my friend is only formed on the basis of his memory (i.e., his background knowledge about me, which doesn't include the results of my introspective exercise). As a result, I might know better than my friend because my belief is formed *also* on the basis of the findings provided by introspection (which I didn't tell my friend) while the belief of my friend is only based on his background knowledge about me (which doesn't include the findings of my introspective exercise). Thus, in this case, to know better is a matter of third-person epistemic self-knowledge: having a true (second-order) belief warranted to a higher degree than my friend's belief about me can possibly be (in that particular situation).

Henceforth, all the arguments and clarifications that were made for the case of memory can be made, *mutatis mutandis*, for the case of introspection as well. Firstly, it is not always the case that I can know better the expressive content of my behaviour in a counterfactual situation than third-person subjects who know me well. For instance, I can exercise my expressive capacity of introspection aloud rather than silently in thought (e.g., saying "I feel terrible just to imagine that my partner is telling me that she is leaving me for another person"), and so, my friend could take into account the results of my introspective exercise to make a judgement about the expressive content of my behaviour in that counterfactual situation. Secondly, the difference between the warrant of my second-order belief and any possible warrant that my friend's belief could have in this particular situation (before I tell him the findings of my introspective exercise) is a difference in *degree* and not in *type*. For both he and me apply a third-person epistemic method (i.e., making judgements about someone's mental states) on the basis of the same type of evidence (i.e., my expressive episodes), although in my case I can use not only the background knowledge about myself but also the new introspective-states provided by introspection. And, thirdly, the fact that third-person self-inspective judgements about the expressive content of my behaviour in a hypothetical situation are a genuine phenomenon can be proved by examples such as the following one. Imagine that my friend asks me how I'd feel in the hypothetical situation of a breakup. I exercise my expressive capacity of introspection silently in thought from the first-person perspective and I start to feel introspective-despair. However, I know from past experiences that when I introspect myself, I tend to express introspective-states whose intentional content points to feelings that are more negative than the feeling that I actually end up having when the hypothetical situation actually takes place. So, rather than answering my friend's question with the first-person self-interpretation "I'd feel terrible", I answer with a third-person self-interpretation or self-inspective judgement on the basis of the evidence about my mental states provided both by introspection and memory (i.e., my introspective-despair and my background knowledge about me); particularly, I answer saying "I think that I would feel depressed and that I would need some time to recover, but it wouldn't be as terrible as it could be".

Therefore, the behavioural-expressivist account can explain the examples described at the beginning of this section claiming that self-interpretations can be either *first-person self-interpretations* or *third-person self-interpretations*. On the one hand, first-person self-interpretations are *expressions* issued on the basis of no evidence about my current mental states. Particularly, first-person self-interpretations of the expressive content of my behaviour from the past are expressive episodes of memory-states (i.e., mental states whose intentional content is about other episodes of mental states that I expressed in the past), and first-person self-interpretations of the expressive content of the behaviour that I would have in a hypothetical situation are expressive episodes of introspective-states (i.e., mental states whose intentional content is about other episodes of mental states that I would express in the relevant hypothetical or counterfactual situation). The fact that first-person self-interpretations are expression has the advantage of avoiding the difficulties of Finkelstein's account to characterize the first-person perspective. For, whereas Finkelstein considers that the first-person perspective is "a broader genus" that includes the species of expression and the species of first-person self-interpretations, the behavioural-expressivist account can characterize the first-person perspective in the following way. Firstly, the first-person itself (i.e., "me" rather than "you") is identified with the broader genus of expression. In turn, the first-person perspective (from which first-person self-interpretations are made) is identified with expressing a first-order mental state on the basis of no evidence about one's own mental states, and the third-person perspective (from which third-person self-interpretations are made) is identified with expressing a second-order belief on the basis of evidence about one's own mental states (i.e., by self-inspection).

On the other hand, the behavioural-expressivist account considers that third-person self-interpretations are judgements about the expressive content of my behaviour in the past or about my behaviour in a hypothetical situation made on the basis of evidence about my mental states. These judgements or third-person self-interpretations are expressive episodes of *second-order beliefs* about my mental states, and so, these judgements or third-person self-interpretations might be instances of third-person epistemic self-knowledge (i.e., true warranted belief). Moreover, the account proposed here avoids the objection raised against Cassam's inferential account because it considers that not all evidence about one's own mental states is inferential evidence. When self-inspecting myself from the third-person perspective, I might use perception, memory and introspection as epistemic sources of evidence about my own mental states. These are epistemic sources of *non-inferential evidence* insofar as they provide a content from which an inference about my mental states can be made, but the content that they provide is not inferred by me from any previous content. Admittedly, perception (i.e., seeing one face in the mirror) is rarely used when one makes judgements about one's own mental states. However, we often make self-inspective judgements on the basis of the expressive episodes of memory-states provided by memory and on the basis of the expressive episodes of introspective-states provided by introspection. Also, against Cassam, the account proposed here considers that all evidence about my mental states is external evidence; however, not all external evidence is perceptual evidence (for I rarely use perception when self-inspecting myself). All evidence about my mental states is external evidence because perception, memory and introspection provide me with different *forms of presentation* of the same reference: my expressive behaviour or expressive episodes of mental states (e.g., my sad facial expression in the mirror, my sad expressive episodes in the past, or my sad expressive episodes in a hypothetical situation).

Moreover, not all the expressive episodes of mental states are presented with the same degree of clarity to perception, memory and introspection. The degree of clarity with which an expressive content of an episode of mental state is presented depends on the vehicles of expression in which it is instantiated and on the conditions in which the relevant epistemic source of evidence (perception, memory or introspection) works (e.g., perceiving enough portion of the pattern of expression and enough context, not being drunk when exercising one's memory or introspection, etc.). An action (i.e., expressive vehicle) that I performed in the past could be presented to my memory in a clearer way (i.e., as a clear episode of the desire to have an ice cream) than to someone else's perception if she doesn't have enough context (e.g., she doesn't see me picking up an ice-cream). However, the clear and aloud utterance "I want an ice cream" (i.e., expressive vehicle) can be presented with the same degree of clarity (i.e., as a clear episode of the desire to have an ice cream) to my memory than to someone else's perception if the appropriate conditions take place (e.g., if there's no too much noise in the kitchen). As a result, third-person avowals (e.g., "I like ice cream a lot") sometimes can be authoritative because they can express second-order beliefs whose warrant is stronger in degree than other people's belief about my mental states can possibly be in that particular situation (e.g., before I tell them how many times I desperately tried to find an ice cream this summer). Thus, when they are true, they might be instances of third-person authoritative self-knowledge in that particular situation.

In this last half of this chapter, it has been argued that neo-expressivist accounts of first-person self-knowledge have in common with epistemic accounts of Transparency that they understand first-person self-knowledge as an epistemic phenomenon (i.e., true warranted belief). Then, it has been argued that the idea of first-person self-knowledge as an epistemic phenomenon is conceptually flawed, and the behavioural-expressivist notion of first-person self-knowledge as an expressive phenomenon (*knowing how*) has been explained. Finally, it has been shown that self-knowledge as an epistemic phenomenon (i.e., true warranted second-order belief) is a third-person phenomenon and that behavioural expressivism can explain third-person epistemic self-knowledge better than other accounts (i.e., Cassam's inferential account).

In the next two chapters, it is going to be argued that the behavioural-expressivist account of Transparency explains self-deception and Moore's paradox better than the accounts currently available in the literature. Since quite a few of the available accounts of self-deception and Moore's paradox share with epistemic accounts of Transparency the idea that first-person avowals are self-ascriptions of mental states, to argue in favour of the behavioural-expressivist account of self-deception and Moore's paradox is to argue against the idea that first-person avowals are self-ascriptions of mental states, and so, against epistemic accounts of Transparency. Indeed, if behavioural expressivism offers the best explanation of self-deception and Moore's paradox, it can be inferred that behavioural expressivism offers the best explanation of Transparency as well because its ideas explain more than the ideas that follow from epistemic accounts of Transparency.

3. Self-Deception

Self-deception is a complex and elusive phenomenon. So, in order to get an idea of what it is, it might be best to see an example of it from the beginning. Let's describe John's case of self-deception:

> John has recently discovered a disturbing spot on the back of his shoulder. As it is on the back of his shoulder, he doesn't know for how long it has been there. So, the possibility that it is malignant and that it wouldn't respond to treatment makes him especially nervous. Since John discovered the disturbing spot, he has been acting oddly in regard to doctors and medical issues. For instance, he avoids talking about the need to have the spot checked by a professional, he refuses to go to the doctor when he gets sick, or he fails to attend check-ups in spite of being mandatory at his company. Furthermore, John has been asking himself about the possibility of an afterlife in spite of not being a religious person and he has enhanced the policy of his health insurance —John lives in a country without universal healthcare— more than would haven been financially responsible. Lydia, his friend, has noticed something odd in John's recent behaviour regarding medical issues. So she asks John: "Do you believe that you might have some kind of illness?". Unexpectedly, John sincerely answers: "Not at all. [I believe that] I am healthy. I'm as fit as a fiddle!".
>
> Lydia knows that John is not lying to her. For Lydia knows that John and his partner have recently decided to have their first child and that John wouldn't have taken that

> decision if he thought that he might have an important illness. However, Lydia is not satisfied with the answer just yet. So, she continues asking: "Then, why do you avoid going to the doctor? Why are you paying so much for your health insurance?". John answers these questions giving reasons that allegedly justify his actions. For instance, sincerely saying "I've been very busy lately and I don't have enough time to go to the doctor" or "It is important to be cautious in life and I can afford the new policy"). However, these alleged reasons look like excuses to Lydia. For Lydia, who knows John's personality and life situation quite well, knows that John has always been up for a beer lately, that he usually drives without fastening his seat belt, and that he had to borrow some money from her to pay the rent. However, Lydia doesn't consider appropriate to keep questioning John for the moment.

John is self-deceived about his health condition because he exhibits the kind of *motivated irrational attitude* towards the fact that p (i.e., that he is healthy) that is characteristic of self-deception. On the one hand, John's attitude is irrational because there is a conflict between what John sincerely *says* (e.g., "[I believe that] I am healthy") and how he *acts* (e.g., avoiding mandatory check-ups or enhancing the policy of his health-insurance more than he should). On the other hand, John's attitude is *motivated* because it is the result of having found a *disturbing* spot on the back of his shoulder. Insofar as John's case intuitively belongs to the *explanandum* of every account of self-deception (at least prima facie), John's case is a *paradigmatic* example of self-deception. Of course, as a result of its *explanans*, a particular account of self-deception can conclude that John's case is not a genuine instance of self-deception[32], but it can only do that at the cost of being in a disadvantage over other competing accounts with a similar degree of simplicity and explanatory power, *plus* the capacity to account for John's case as an example of self-deception.

It is useful to distinguish between 1) the *cause* of self-deception, 2) the *process* by which self-deception is generated, and 3) the *psychological state* (i.e., the mental state or the cluster of mental states) in which subjects are once they become self-deceived[33]. Regarding 1)

[32] For instance, Lynch (2012) would consider John's case an example of *escapism* (i.e., avoiding an unpleasant truth that one knows about) instead of an example of self-deception (i.e., not believing an unpleasant truth because one is biasedly deceived). For Lynch's account is not powerful enough to offer a homogeneous explanation of both John-like cases (in which the subject *apparently* has two contradictory beliefs) and garden-variety cases of self-deception (in which the subject *apparently* has a single —motivated— belief) at the same time.

[33] I borrow this terminology from Van Leeuwen (2007a, 2007b), although I have adapted it for my purposes.

the cause of self-deception, there is a widespread agreement about the idea that self-deception is a motivated phenomenon[34]. Unlike in cases of unmotivated mistakes, people become self-deceived only about issues that matter to them because self-deception is always caused by a motivational state (i.e., the desire that p, the desire to *believe* that p or an emotion about p) whose intentional content is related to the intentional content of self-deception in some way. The relation between the intentional content of the motivational state and the intentional content of self-deception can be *straight* or *twisted*. In straight cases of self-deception, subjects are self-deceived about p because they have a positive motivational state towards p (i.e. they want p or p makes them happy). For instance, I might be self-deceived about the fact that I am handsome because I want to be handsome. In twisted cases of self-deception, by contrast, subjects are self-deceived about p because they have a negative motivational state towards p (i.e. they don't want p or p makes them anxious). For instance, a jealous husband might be self-deceived about the fact that his wife is unfaithful because he is afraid of the possibility that his wife is unfaithful[35].

Unfortunately, the agreement in the literature stops beyond the idea that 1) the cause of self-deception is a motivational state. 2) The particularities of the process by which self-deception is generated and 3) what is precisely the psychological state (i.e., the mental state or cluster of mental states) in which subjects are once they become self-deceived are both widely discussed issues. The state of the art can be accurately described by characterizing the disagreement between the accounts that identify self-deception with 2) a process that generates a false or unwarranted belief (henceforth, *procedural accounts*) and the accounts that identify self-deception with 3) a *sui generis* psychological state (henceforth, *psychological-state accounts*).

Firstly, procedural accounts of self-deception think that self-deception is identical to a certain kind of *process* that generates a false or unwarranted belief (which is supposed to be the product of self-deception). There are three types of procedural accounts of self-deception: *motivationalist accounts* (Barnes, 1997; Johnston, 1988; Lauria et al., 2016; Lazar, 1999; Lynch, 2012; Mele, 1999, 2001, 2008; Nelkin, 2002, 2012; Szabados, 1974, 1985; Van Leeuwen, 2007a, 2007b), *intentionalist accounts* (Bermudez, 2000, 1997; Davidson, 2004, 1985, 1998; Demos, 1960; Foss, 1980; Pears, 1984, 1991; Rorty, 1988; Sorensen, 1985;

[34] Patten (2003) is the only author (that I know of) who explicitly argues that self-deception is not motivated.

[35] See Barnes (1997, Ch. 3), Mele (1999; 2001, Ch. 5), Nelkin (2002) and Scott-Kakures (2001, 2002) for different accounts of straight and twisted cases of self-deception.

Talbott, 1995; Trivers, 2011) and *epistemic accounts* (Fernández, 2012, 2013; Funkhouser, 2005, 2016; Holton, 2001; Patten, 2003; Sandford, 1988; Scott-Kakures, 1996, 1997, 2002). Motivationalist accounts claim that self-deception is identical to a motivated bias that operates in the deliberative process by which self-deceivers acquire and retain their beliefs in the teeth of evidence to the contrary and that that motivated bias is *enough* for self-deception to take place. By contrast, intentionalist accounts and epistemic accounts think that, even if self-deception involves some kind of motivated bias, the existence of self-deception requires some *additional* condition. On the one hand, intentionalist accounts claim that self-deception also requires the intention of the subject to deceive herself by forming a belief that she considers to be false. On the other hand, epistemic accounts claim that self-deception also requires an epistemic failure in the process of first-person self-knowledge for which the subject is epistemically responsible.

Secondly, psychological-state accounts of self-deception identify self-deception with the *sui generis* psychological state (i.e., a *sui generis* mental state or a *sui generis* cluster of mental states) in which subjects are supposed to be once they become self-deceived and not with the process that has led them to be in that psychological state. According to these accounts, regardless of the process by which self-deception is produced (which would be one involving a motivated bias), self-deception is a *sui generis* psychological state of variable complexity. Thus, it has been argued that self-deception consists in *pretending* that p is the case while one doesn't believe so (Gendler, 2010, Ch. 8), in the *sincere avowal* that p without the corresponding belief that p (Audi, 1982, 1988, 1989, 1997; Rey, 1988), or in avoidance of the distressing recurrent thought that not-p when the subject believes that not-p but desires that p (Bach, 1981).

In this chapter, it is going to be argued that from the behavioural-expressivist account of Transparency follows the best account of self-deception currently available in the literature. The chapter is going to have the following structure. Firstly, the desiderata of a good account of self-deception are going to be glossed out. Secondly, the procedural accounts of self-deception (i.e., intentionalist accounts, motivationalist accounts and epistemic accounts) are going to be explicated and criticized. Thirdly, the most relevant psychological-state accounts are going to be explicated and criticized. Fourthly, it is going to be argued that there are two concepts of truth (i.e., a relational and a non-relational concept of truth) that follow from behavioural expressivism and that unconscious mental states can be appropriately characterized by claiming that they are necessarily false in the non-relational sense. Fifthly, a behavioural-expressivist account of self-deception is going to be proposed and developed, according to which self-deception is a *sui generis* mental state (different from beliefs and from any other conscious attitude) that belongs to the class of unconscious mental states. Sixthly, it is going to be argued that the proposed account of self-deception is the only account currently available in the literature that meets all the desiderata of a good account of self-deception. And, finally, it is going to be shown that the proposed behavioural-expressivist account of self-deception has the additional virtue of distinguishing in an appropriate way different phenomena which are considered to be closely related: wishful thinking, self-deception and delusion.

3.1 The desiderata of self-deception

There could be cases of self-deception in which subjects sincerely issue first-person avowals that make explicit the mental state of desire (e.g., "I want to get married"), cases in which subjects issue first-person avowals that make explicit the mental state of intention (e.g., "I intend to join the gym this month") and cases in which subjects issue first-person avowals that make explicit the mental state of belief (e.g., "I believe that I am healthy"). John's case is going to be used in the remainder of this chapter as a *paradigmatic* example of a belief-type case of self-deception, meaning that it is going to be used as a case of self-deception from which glossing out the key features of the phenomenon of self-deception and from which giving rise to the list of desiderata that every account of self-deception should be able to meet to explain self-deception in an appropriate way. Then, let's see the different intuitions or desiderata about self-deception that every good account of self-deception should be able to explain:

1) It should be able to explain the *irrational conflict* between what the subject *says* and how he *acts* (e.g., Audi, 1982; Davidson, 2004; Demos, 1960; Fernández, 2012, 2013; Foss, 1980; Funkhouser, 2005; Funkhouser & Barret, 2016).

   On the one hand, it is not enough for John to be self-deceived that he systematically avoids going to the doctor when he gets a cold or that he decides to enhance the policy of his health insurance more than he should. For this kind of behaviour could be manifested by a non-self-deceived subject who simply happens to manage some aspects of his life irresponsibly. On the other hand, it is not enough for John to be self-deceived that he sincerely says "I believe I'm healthy", without manifesting any conflicting behaviour, after having seen the spot on the back of his shoulder. For this kind of linguistic behaviour (i.e., saying "I believe that I am healthy) could be manifested by a non-self-deceived subject who actually believes that the spot is just a regular mole and that he is perfectly healthy. Therefore, an irrational conflict between what the subject sincerely says (e.g., "I believe that I am healthy", "I've been busy lately", "I like to be cautious and I can afford it") and how he acts (e.g., avoiding doctors' appointments, enhancing his health insurance more than he should, being always up for a beer, driving without fastening his seatbelt, etc.) is necessary for self-deception to occur.

2) It should be able to explain why subjects seem to have some kind of *unconscious* mental state or to suffer from some kind of *lack of self-knowledge* when they are self-deceived (e.g., Audi, 1982; Bermudez, 1997; Demos, 1960; Fernández, 2012, 2013; Funkhouser & Barret, 2016; Holton, 2001; Lazar, 1999; Pears, 1984; Talbott, 1997). Indeed, the existence of some unconscious mental state or of some kind of lack of self-knowledge seems to be needed in order to explain John's conflicting behaviour between what he *sincerely* says (e.g., "I believe that I am healthy") and how he acts (e.g., avoiding check-ups, enhancing his health insurance, etc.). For, without positing some unconscious mental state or some kind of lack of self-knowledge, it would be more natural to depict John as lying to Lydia than as being self-deceived about his health condition. Also, the presence of such unconscious mental state or of such lack of self-knowledge should be able to explain why subjects often say things like "I was fooling myself" or "Deep down, I knew the truth all along" after overcoming self-deception.

3) It should be able to explain the high degree of *persistence* against criticism based on evidence that self-deception has (e.g., Fernández, 2013, pp. 191-192; Lazar, 1999, p. 267) because of the *rationalizations* of the subject.

When people are mistaken about a motivationally neutral issue and the error is pointed out to them adducing good evidence, people often correct their mistake changing their mind about the issue. However, self-deception is not like an unmotivated mistake in that regard. On the one hand, it is easy to imagine John twisting the evidence about his health condition to defend his alleged belief that he is healthy if Lydia pointed out to him (adducing good evidence) that the spot on the back of his shoulder is serious enough to have it checked. On the other hand, it is easy to imagine John making up (sincere) excuses in an attempt to justify his irrational behaviour if Lydia pointed out to him (adducing good evidence) that he has been acting irrationally in regard to his health condition. This kind of self-deceptive behaviour is usually called "rationalization".

Thus, the persistence of self-deception is due to the following two kinds of rationalizations. On the one hand, self-deceivers tend to *rationalize the evidence* about the issue that they are self-deceived about (e.g., Audi, 1997, Ch. 6; Bach, 1981; Barnes, 1997; Davidson, 2004; Johnston, 1988; Lauria et al., 2016; Lazar, 1999; Lynch, 2012; Mele, 1999, 2001, 2008; Nelkin, 2002, 2012; Pears, 1984; Szabados, 1974, 1985; Talbott, 1995; Van Leeuwen, 2007a, 2007b) because they tend to have a motivated bias in favour of the fact that they allegedly believe that it is the case. So, when evidence against the fact that they say to believe is given to them, self-deceivers tend to assess the evidence in a twisted and biased way, preserving so their self-deceptive state. For instance, John might not take it as seriously when his partner tells him that the spot on the back of his shoulder looks suspicious and that he should have it checked by a doctor than when a friend tells him that it would be unlikely that the spot is something malignant at his age. On the other hand, self-deceivers tend to *rationalize their conflicting behaviour* (e.g., Audi, 1997, Ch. 6; Bach, 1981; Funkhouser & Barret, 2016; Patten, 2003; Sandford, 1988; Szabados, 1985) because they tend to make up (sincere) excuses in an attempt to justify their actions when evidence against their rationality is given to them, preserving so their self-deceptive state. They are *sincere* excuses because, in spite of being adduced by self-deceivers without the intention to lie, they are not the real reasons that explain why their actions were performed. For instance, when Lydia asks John why he avoids going to the doctor or why he is enhancing his health insurance, he sincerely answers that he has been very busy lately (even if he was always up for a beer) and that he thinks it is important in life to take precautions just in case (even if he usually drives without fastening his seat belt). In a broader context of discussion than the study of self-deception, this phenomenon has been called *confabulation* (e.g., Hirstein, 2005; Keeling, 2018; Sullivan-Bissett, 2014).

4) It should be able to explain why subjects are *epistemically responsible* for being self-deceived so that they are held accountable for having such an irrational state (e.g., Bach, 1981; Demos, 1960; Fernández, 2012, 2013; Foss, 1980; Holton, 2001; Johnston, 1988; Nelkin, 2012)[36].

There are cases in which subjects manifest episodes of conflicting behaviour that are similar to John's example but aren't intuitively cases of self-deception because the subjects are not epistemically responsible for having the irrational state. For instance, it has been reported that some schizophrenic patients sincerely say that they are afraid of their caregivers because they are trying to kill them, but they willingly keep eating the food that their caregivers provide them with[37]. We don't think, though, that those schizophrenic patients suffer from self-deception in spite of manifesting episodes of conflicting behaviour (i.e., between what they sincerely say and how they act) analogous to cases of self-deception because we don't think that they are epistemically responsible for thinking that their caregivers are trying to kill them. By contrast, it is easy to imagine Lydia criticizing John and holding him epistemically accountable for twisting the evidence and for avoiding going to the doctor with (sincere) excuses. Why do we consider self-deceived subjects, but not schizophrenic subjects, epistemically responsible for their irrational states?

It has been argued (McHugh, 2013) that epistemic responsibility in regard to belief has two necessary conditions: that the subject is the agent of the belief (*doxastic agency*) and that the belief can be modified on the basis of reasons (*reason-*

[36] Usually, the subject's responsibility for being self-deceived is discussed focusing on *moral* responsibility. However, I focus on *epistemic* responsibility instead because I think that self-deceivers are *always* epistemically irresponsible, but (against Levy, 2004) I think that self-deceivers are morally irresponsible only *sometimes*. For instance, morally speaking, it is not the same to be self-deceived in thinking that your son is going to be found alive after getting lost in the woods (for nobody is harmed by the actions involved in that self-deception) than to be self-deceived in thinking that you are not a racist when you are one as a matter of fact (for others are harmed by the actions that that self-deception involves or makes possible).

[37] I found this case in Fernández (2012, p. 382; 2013, p. 185).

responsiveness). It can be doubted whether a schizophrenic patient is the doxastic agent of the thought that his caregivers are trying to kill him (although it is difficult to deny that subjects with schizophrenia generally engage in deliberations about facts of the world as any other subject). However, it seems clear that the schizophrenic patient's irrational thought is not reason-responsive or reason-sensitive (so that he is not epistemically responsible for having that thought). The idea of reason-responsiveness can be explained using the concept of *threshold*. Self-deception is a reason-sensitive psychological state because there is a threshold of evidence that self-deceivers can reach in order to overcome self-deception. For instance, if John finds out that the suspicious spot on the back of his shoulder is growing and growing so that the amount of evidence pointing to the fact that it is malignant become overwhelming, John will overcome self-deception eventually and he will start to believe that he might be ill. Of course, the threshold of evidence that needs to be reached in order to overcome self-deception is much higher than the threshold of evidence that needs to be reached in order to overcome an unmotivated mistake, but there is still a threshold of evidence that is possible to reach in order to overcome self-deception. Since self-deception is reason-responsive or reason-sensitive in this sense, self-deceivers are epistemically responsible for being self-deceived. By contrast, the irrational state of the schizophrenic patients is not reason-responsive or reason-sensitive in this sense because there isn't any threshold of evidence that they can reach in order to overcome the thought that their caregivers are trying to kill them. No matter how much evidence they could have available against the fact that their caregivers are trying to kill them; if they finally change their mind, it won't be because of that evidence. Since the irrational psychological state of the schizophrenic patients is not reason-responsive, they are not epistemically responsible for having that irrational psychological state. That's why schizophrenia, unlike self-deception, is treated with specific medicines and not with psychoanalysis (or something of the like).

Therefore, every good account of self-deception should be able to explain 1) the irrational conflict between what self-deceivers say and how they act, 2) why self-deceivers seem to have some kind of unconscious mental state or to suffer from lack of self-knowledge,

3) how self-deceivers rationalize both the evidence and their actions to preserve their self-deception and 4) why self-deceivers are epistemically responsible for being self-deceived. So, let’s start seeing how the procedural accounts of self-deception deal with these desiderata.

3.2. The procedural accounts of self-deception

This section will discuss the three different kinds of procedural accounts of self-deception (i.e., which identify self-deception with the process that generates a false or unwarranted belief) available in the literature: intentionalist accounts, motivationalist accounts and epistemic accounts. It will be concluded that none of them manages to explain self-deception in an appropriate way because they fail to account for at least one of the desiderata of self-deception.

3.2.1 Intentionalist accounts

Intentionalist accounts claim that self-deception is analogous to interpersonal deception, being the difference that in cases of self-deception the deceiver and the deceived are one and the same person. The process of self-deception is thought to go as follows. Firstly, John truly believes that he is ill because of the spot on the back of his shoulder. Secondly, he forms the intention to deceive himself into thinking that he is healthy (to avoid psychological pain or for whatever motivational reason). Finally, the intention is fulfilled and John ends up believing at the same time both that he is ill (the original true belief) and that he is healthy (the intentionally self-caused false belief). Therefore, the irrational conflict between what the self-deceived subject sincerely says and how he acts is due to the fact that he has these two contradictory beliefs. On the one hand, John enhances his health insurance more than he should because he genuinely believes that he is ill and he avoids going to the doctor because he is convincing himself of the fact that he is healthy (what involves avoiding collecting evidence from places where it could be established that he is ill). On the other hand, John sincerely says

"I believe I'm healthy" and has decided to have a child because, as a result of his *successful* attempt to deceive himself, he also believes that he is healthy now.

However, intentionalist accounts face two significant problems. They fall into what Mele calls the *static* and *dynamic* puzzles (2001, Ch. 3). On the one hand, they fall into the static puzzle because it seems highly implausible that a normal subject, without impaired cognitive faculties, could believe a straight contradiction. On the other hand, they fall into the dynamic puzzle because it seems highly implausible that an intention of the subject to deceive herself, forming a belief that the own subject considers false, could succeed in eliciting the intended belief.

Intentionalist accounts have tried to overcome these perplexities following different strategies to explain away the static and the dynamic puzzles[38]. Firstly, some authors have followed the *attention strategy* (Demos, 1960; Foss, 1980). According to the attention strategy, there is nothing puzzling in thinking that self-deceivers have a pair of contradictory beliefs (static puzzle) because what happens is that self-deceivers don't pay attention to both beliefs at the same time, and so, they are not aware of the contradiction. Secondly, other authors have followed the *temporal strategy* (Bermúdez, 2000; Sorensen, 1985). According to the temporal strategy, self-deception is a temporal process in which subjects start believing that p and end up believing that not-p, without ever believing at the same time both p and not-p (that is, without ever believing a contradiction —static puzzle—). Finally, some authors have followed the *division strategy* (Davidson, 2004, 1998; Pears, 1984, 1991; Rorty, 1972, 1988). According to the division strategy, the mind of self-deceived subjects is somehow divided into two parts so that there is no single part which plays the role of the deceiver and the deceived at the same time (avoiding both the static and the dynamic puzzles). One part of the mind contains the true belief that p is the case and the intention to deceive oneself about p; and the other part of the mind contains the contradictory belief that not-p, which has been produced by the subject's intention to deceive herself.

However, none of these strategies works well enough to guarantee that some progress has been made. Leaving aside the fact that the attention and the temporal strategies are useless to tackle the *dynamic* puzzle at all (which already is an important weakness), I will discuss the main difficulties that these three strategies face to explain the *static* puzzle. Firstly, the main

[38] For a full discussion about the strategies of intentionalist accounts to overcome the puzzles, see Fernández, 2013, pp. 188-193.

problem of the attention strategy is that it makes unexplainable the *persistence* of self-deception against criticism based on evidence. If the solution of the static puzzle were that self-deceivers have two contradictory beliefs but they don't pay attention to both at the same time, self-deceivers would naturally overcome their irrational state just by being warned of the fact that they have two contradictory beliefs. Thus, self-deception would lack the degree of persistence that it has compared to regular mistakes. For instance, John would easily overcome self-deception just by letting him know (with some good evidence in support) that there is an irrational conflict between what he says and how he acts, as if he had the contradictory beliefs that he is healthy and that he is ill at the same time. For, at the moment in which he is being told, it would be impossible for him to avoid paying attention to the two contradictory beliefs at the same time.

Secondly, the temporal strategy fails to account for the conflicting behaviour characteristic of self-deception. The temporal strategy explains away the static puzzle claiming that self-deceivers believe that p at $t_{x\text{-}y}$ and that not-p at $t_{y\text{-}z}$, without ever believing p and not-p at the same time. However, this conception of self-deception is mistaken. The conflicting behaviour manifested by self-deceivers is not such that in $t_{x\text{-}y}$ they act *only* as if they believed that p and in $t_{y\text{-}z}$ they act *only* as if they believed that not-p (so that their behaviour conflicts between $t_{x\text{-}y}$ and $t_{y\text{-}z}$, but not *within* $t_{x\text{-}y}$ or *within* $t_{y\text{-}z}$). By contrast, the behaviour of self-deceivers is such that they manifest the conflict between what they say and how they act all over the time that they are self-deceived ($t_{x\text{-}z}$). For instance, John says "I believe that I am healthy" (as if he believed that he is healthy) while enhancing his health insurance (as if he believed that he is ill), and John decides to have a child (as if he believed that he is healthy) while skipping his check-up (as if he believed that he is ill). Then, it is false that John acts *only* as if he believed that he is ill at $t_{x\text{-}y}$, and *only* as if he believed that he is healthy at $t_{y\text{-}z}$; by contrast, he acts in a conflicting way all over $t_{x\text{-}z}$.

Finally, the division strategy faces the problem of giving rise to greater perplexities than the original static and dynamic puzzles themselves[39]. The division strategy faces the following dilemma: if it takes the compartmentalization of the mind *seriously*, it solves the puzzles at the cost of endorsing a highly implausible view of the mind; and if it takes the compartmentalization of the mind *metaphorically*, it doesn't solve the puzzles at all. On the one hand, if the division strategy takes the compartmentalization of the mind seriously and self-

[39] For a full discussion on the paradoxes produced by the division strategy, see Johnson, 1988 p. 79-86.

deceivers are depicted as having two different subsystems (each one with their own beliefs, intentions and desires inaccessible to the other), then the division strategy overcomes the static and dynamic puzzles, but it ends up endorsing a very implausible view of the self-deceivers' mind: as if self-deceivers hosted two different *selves* within the same person. On the other hand, if the division strategy takes the compartmentalization of the mind metaphorically and it claims that there are different subsystems in the self-deceivers' mind but that they don't constitute independent selves because they all have access to all the belief, desires and intentions of the others, then the static and dynamic puzzles remain unsolved insofar as the following questions remain without answer: How can beliefs, intentions and desires be attributed to one of the subsystems of the self-deceived subject without attributing them also to the other subsystems? How can the deceived subsystem fall into the traps of the deceiver subsystem if the former has access to what the latter is up to?

Therefore, none of the strategies of intentionalist accounts to overcome the static and dynamic puzzles works well enough to avoid failing to explain some key features of self-deception (i.e., the persistency of self-deception and the characteristic conflicting behaviour) nor well enough to avoid giving rise to greater perplexities than the original static and dynamic puzzles themselves (i.e., the idea of two selves within the same person).

3.2.2 Motivationalist accounts

Given the failure of intentionalist accounts to offer a plausible account of self-deception, motivationalist accounts emerge as a natural alternative to explain self-deception without puzzles. Motivationalist accounts think that self-deception doesn't require either a pair of contradictory beliefs or the subject's intention to deceive oneself, and so, they avoid the static and dynamic puzzles from the outset. Different motivationalist accounts have given different conditions of self-deception[40], but all of them share the same core idea. A subject is self-deceived about p when the following conditions are met:

[40] One of the most prominent motivationalist authors is A. Mele. He famously offers four jointly sufficient (but not necessary) conditions of self-deception (Mele, 2001, pp. 50-51). Here they are:

1. The belief that p which S acquires is false.

1) It is the case that not-p[41], and so, it is the case that the total amount of evidence objectively available supports the fact that not-p.
2) The total amount of evidence available to the subject, or that *should have been* available to her[42], supports the true belief that not-p according to her own *epistemic norms*[43].
3) The subject *collects* and *assesses* the evidence in a way that goes against her own epistemic norms because of the influence of a motivational state (e.g., a desire, an emotion, etc.).
4) The subject ends up forming the false belief that p as a consequence of the fact that she (unintentionally) collects and assesses the available evidence in a motivationally biased way.

Therefore, it follows from motivationalist accounts that John is self-deceived about his health condition because it is a fact that he is not healthy, it is a fact that the evidence available supports the idea that he is not healthy according to John's epistemic norms, but he ends up forming the false belief that he is healthy because he (unintentionally) collects and assesses the evidence in a motivationally biased way (i.e., against his own epistemic norms). There are different ways in which the subject's collection and assessment of the evidence about p can be implicitly biased. Mele gives four examples of ways in which a motivational state can bias the subject's collection and assessment of the evidence without the subject's intention to deceive herself or to manipulate the evidence (Mele, 2001, pp. 25-27):

2. S treats data relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way.
3. This biased treatment is a nondeviant cause of S's acquiring the belief that p.
4. The body of data possessed by S at the time provides greater warrant for not-p than for p.

[41] Van Leeuwen (2007a, 2007b) is the only motivationalist author that I know of who doesn't explicitly mention this condition of self-deception.

[42] I introduce this qualification to take into account those cases in which the evidence available to the subject supports the false belief that not-p only because the subject has collected the evidence in a biased way against her own epistemic norms. For instance, having handy a set of evidence that points to the true belief that p, but refusing to collect that set of evidence because of the influence of a motivational bias in favour of the belief that not-p. Nelkin (2002, pp. 393-398) also makes this point.

[43] This qualification is introduced by Van Leeuwen (2007a, 2007b). It is an important qualification to rule out cases in which the evidence objectively supports the true belief that p from an epistemically rational perspective, but the subject ends up forming the false belief that not-p because of her poor epistemic norms and not because of a motivational bias. For instance, a subject who forms the false belief that it is going to rain, in spite of the fact that the weather forecast says the opposite, because she believes that she can know whether it is going to rain by flipping a coin (epistemic incompetence).

1) *Negative misinterpretation*: the subject takes evidence that counts against not-p according to her own epistemic norms as not counting against not-p. John finds a *prima facie* suspicious spot on the back of his shoulder and, after a motivationally biased deliberation, he concludes that the spot is actually nothing more than a regular mole. Had John not been motivationally biased (e.g., by the desire to be healthy), he would have concluded that the spot deserves a check-up. For this is the belief that the evidence available to John actually supports, according to his own epistemic norms.

2) *Positive misinterpretation*: the subject takes evidence that counts against not-p according to her own epistemic norms as counting in favour of not-p. Let's suppose, for instance, that John loses 20 pounds of weight in a few weeks. According to John's criteria, to lose 20 pounds of weight in a few weeks is *prima facie* a symptom of the fact that something wrong is going on. However, after a motivationally biased deliberation, John concludes that his particular weight loss is due to the fact that he has eaten healthier lately (despite not having much evidence to support this idea). Had John not been motivationally biased, he would have concluded that something must be wrong in his body. For this is the belief that the evidence available to John actually supports according to his own epistemic norms.

3) *Selective focusing or attending*: the subject fails to pay attention to evidence that seems to count against not-p, focusing instead on evidence that seems to count in favour of not-p, against her own epistemic norms. For instance, against his own epistemic norms, John avoids talking about medical issues because that reminds him of the spot on the back of his shoulder, which counts in favour of the fact that he might be ill; and he tends to focus instead on how strong he usually feels when he wakes up early to go to work. Had John not been motivationally biased, he wouldn't have avoided any conversation about medical issues.

4) *Selective evidence-gathering*: the subject tends to gather new evidence where it is likely to find evidence in favour of not-p and avoids gathering new evidence from

> sources that are likely to provide her with evidence against not-p. In doing that, she goes against her own epistemic norms regarding both the different sources of evidence that are supposed to be taken into account and the quantity of evidence that is supposed to be gathered from them before reaching a conclusion. For instance, John avoids going to the doctor because the doctor could tell him what he doesn't want to hear, in spite of the fact that reaching a conclusion without having the spot checked by a professional goes against his epistemic norms. However, John would pleasantly listen to someone talking about how smooth everything usually works in your body at his age. Had John not been motivationally biased, he would have sought the opinion of a doctor.

These biased ways of collection and assessment of evidence could be considered the causal effect of a more basic *affective mechanism* of distress-and-relief that is triggered by the conjunction of a *motivational state* (e.g., the desire that p, anxiety of p, etc.) and the subject's *suspicion* that p is the case (Mele, 2001, pp. 71–72, 79–80; Van Leeuwen, 2007a, p. 427). Let's see how this distress/relief mechanism, plus a suspicion (e.g., the suspicion that the spot is something serious) and a motivational state (e.g., the desire to be healthy), can explain John's biases without the need of positing that he has the unconscious intention to deceive himself (against Talbott's objection to motivationalist accounts —1995—) and without the need of positing that the subject has unconscious knowledge of the truth (against Funkhouser's & Barrett's objection to motivationalist accounts —2016—).

Suppose that it is a fact that John is ill and that the evidence available to John supports the true belief that the spot is likely malignant according to John's epistemic norms. When John found the spot on the back of his shoulder, it looked (*prima facie*) suspicious to him. Thus, he started to deliberate about its shape, colour and size (evidence) in order to find out whether it is malignant or just a regular mole. Since John really wants to be healthy, he finds *relief* in the evidence that can be reinterpreted as pointing to the fact that it is a regular mole and *distress* in the evidence that seems to point to the fact that it is malignant. Therefore, John tends to avoid paying attention to the evidence that cannot be reinterpreted as pointing to the fact that the spot is just a regular mole (*selective focusing or attending*) and he tends to reinterpret the evidence pointing to the fact that it is malignant (*positive misinterpretation*). As a result, he ends up biasedly and falsely believing that the spot is just a regular mole (*negative misinterpretation*) against his own epistemic norms, and so, that there is nothing to suspect about it anymore.

Furthermore, once John has concluded the deliberation forming the false belief that it is a regular mole, he avoids talking about the spot with his partner or looking at it in the mirror (*selective evidence-gathering*) because he finds himself uneasy when paying attention to it (*selective focusing or attending*). After all, paying attention to it could lead him to perceive the spot as suspicious again, and so, to reopen the deliberation about whether it is malignant or not.

However, motivationalist accounts fail to explain the first desideratum of self-deception (i.e., the conflict between what the subject says and does). The problem for motivationalist accounts is that not all the behaviour characteristic of self-deceivers is behaviour that affects the subject's gathering and assessment of the evidence; self-deceivers express their self-deception in non-biasing behaviour as well. As a result, motivationalist accounts cannot explain the *totality* of John's conflicting behaviour. They can explain why John says "I believe that I am healthy" (i.e., John is considered to have the false and unwarranted belief that he is healthy), why John avoids doctors' appointments (i.e., *selective evidence-gathering*) or why John avoids talking about medical issues (i.e., *selective focusing or attending*). But they cannot explain why he enhanced his health insurance more than it would have been financially responsible or why he is suddenly interested in the possibility of an afterlife in spite of not being a religious person because this kind of behaviour doesn't have to do either with an implicit bias or with the belief that he is healthy.

To answer this objection, motivationalist accounts only have left two theoretical resources to explain the totality of John's conflicting behaviour. On the one hand, they could claim that, on top of believing that he is healthy (sincerely saying "I believe that I am healthy"), John also believes that he is not healthy. This certainly would explain why he enhanced his health insurance and why he suddenly got interested in religious topics, but it can only do that at the cost of triggering the *static puzzle* (i.e., attributing him two contradictory beliefs), something that motivationalist accounts wanted to avoid from the very beginning. On the other hand, motivationalist accounts could claim that John enhanced his health insurance or got interested in the afterlife because he *suspects* that he is not healthy. After all, both a suspicion (e.g., the suspicion that he is not healthy) and a motivational state (e.g., the desire to be healthy) are on the basis of John's case of self-deception. However, a suspicion doesn't seem enough to explain why John enhanced his health insurance more than he should or why he suddenly got interested in religious topics for the following reasons. Firstly, if this behaviour were explained because of such suspicion, John should be able to tell that he enhanced his health insurance or that he became interested in religious topics because he suspects that he is not healthy.

Secondly, the suspicion that one is not healthy (as opposed to the belief that one is not healthy) seems like the kind of mental state that can lead a subject to start an investigation to find out the truth, but it doesn't seem like the kind of mental state that can lead a subject to act as if he believed that he is ill (e.g., enhancing his health insurance more than it would have been financially responsible). And, thirdly, not only does John not say that he suspects that he is ill, but he is supposed to believe quite the opposite: he is supposed to have the biased and false belief that he is healthy. Then, motivationalist accounts don't seem to be able to explain the totality of the conflicting behaviour characteristic of self-deception because such behaviour includes more than linguistic expressions of the biased belief that p and more than biasing actions of the subject's gathering and assessment of the evidence; it also includes non-biasing behaviour unrelated with the biased belief that p.

3.2.3 Epistemic accounts

Epistemic accounts claim that self-deception is the result of the subject's lack of self-knowledge due to an epistemic error in the first-person process responsible for self-knowledge. There are two kinds of epistemic accounts of self-deception (Fernández, 2012): error-about-mental-state accounts and error-about-justification accounts. On the one hand, error-about-mental-state accounts (Fernández, 2012, 2013; Holton, 2001; Funkouser, 2005) claim that the failure of self-knowledge characteristic of self-deception has to do with the fact that the subject forms a false second-order belief (e.g., the false second-order belief that one believes that one is healthy) about one of her first-order mental states (e.g., the first-order belief that one is ill). On the other hand, error-about-justification accounts (Patten, 2003; Sanford, 1988; Scott-Kakures, 1997, 2002) claim that the failure of self-knowledge characteristic of self-deception has to do with the fact that the subject overestimates the influence of one of her attitudes (e.g., the belief that one is busy) in bringing about either another mental state (e.g., the belief that one shouldn't go to the doctor this time) or an action (e.g., the action of skipping the doctor's appointment once again), so that the subject lacks first-person self-knowledge of the *real* reasons that justify her belief (e.g., the belief that one shouldn't go to the doctor this time) or of the *real* reasons that justify her action (e.g., skipping the doctor's appointment).

However, in spite of its differences, error-about-justification accounts seem to collapse into a particular *kind* of error-about-mental-state accounts. Since the process that epistemic accounts consider to be responsible for first-person self-knowledge is supposed to involve a second-order belief (i.e., a belief about one's own mental states), it is natural to understand the failure of self-knowledge proposed by error-about-justification accounts as involving a *false* second-order belief about the reasons that justify one's own actions (e.g., skipping doctor's appointments) or beliefs (e.g., thinking that one is too busy to go to the doctor). Therefore, both error-about-mental-state accounts and error-about-justification accounts share the common idea that self-deception arises because of an epistemic error of the subject in the process of first-person self-knowledge that causes a *false* second-order belief. Since Fernández's error-about-mental-state account of self-deception is the only epistemic account that, in addition to claiming that self-deception is the result of an epistemic error in the process of first-person self-knowledge, provides a full account of first-person self-knowledge itself, the remainder of this section is going to focus on discussing Fernández's account as a paradigmatic epistemic account of self-deception. Hopefully, the discussion will be relevant to understand the virtues and limitations of the different types of epistemic accounts of self-deception in general.

According to Fernández, self-deception occurs because of an epistemic error of the subject in the *bypass* procedure responsible for both Transparency and first-person self-knowledge. As a reminder, the bypass procedure describes that, in order to achieve first-person self-knowledge, subjects form their second-order beliefs (e.g., "I believe that it is raining") on the basis of the same epistemic grounds (e.g., that I have the appearance of rain when I look through the window) on which they have formed their corresponding first-order beliefs (e.g., "It is raining"). What happens in cases of self-deception is that the subject ends up with a false second-order belief because of an epistemic failure in the bypass procedure due to her own epistemic negligence at following the procedure. Let's assume that John has what *he takes* as appropriate ground (i.e.: the spot on the back of his shoulder) to believe that he is ill (first-order belief), and so, he forms the first-order belief that he is ill (which explains why he acts as if he were ill: enhancing his health insurance, avoiding doctors because of anxiety, etc.). If John has what he takes as appropriate grounds to believe that he is ill (first-order belief), then, according to the bypass procedure, he has appropriate grounds to believe that he believes that he is ill (second-order belief). However, instead of forming the grounded second-order belief that he believes that he is ill, he fails at following the bypass procedure and forms the ungrounded second-order belief that he believes that he is healthy (which explains his verbal behaviour: "I believe I'm healthy"). Thus, John is self-deceived due to his lack of first-person self-knowledge: he (falsely) believes that he has a first-order belief that he actually hasn't, namely, the first-order belief that he is healthy.

Epistemic accounts of self-deception have some important virtues. Firstly, epistemic accounts of self-deception explain the irrational conflict between what self-deceivers sincerely say and how they act (first desideratum) because the false second-order belief (e.g., the second-order belief that one believes that one is healthy) is supposed to explain the self-deceivers' verbal behaviour (e.g., "I believe that I am healthy") and the first-order mental state (e.g., the first-order belief that one is ill) is supposed to explain the self-deceivers' inconsistent behaviour (e.g., enhancing their health insurance, getting suddenly interested in religious things, avoiding doctors' appointments or talks about medical issues, etc.). Since the false second-order belief that I believe that not-p *doesn't contradict* the first-order belief that p (i.e., they have different contents: "I believe that not-p" and "p"), the static puzzle doesn't arise here. Since the false second-order belief that I believe that not-p is the result of an *unintentional* epistemic error of the subject in the process responsible for first-person self-knowledge, the dynamic puzzle doesn't arise either. Secondly, epistemic accounts of self-deception explain the existence of an unconscious mental state (i.e., the first-order mental state) or of lack of self-knowledge (second desideratum) because the subject has a false second-order belief. As a result, it is explained why subjects say things like "I was fooling myself" or "Deep down, I knew it all along": they had the first-order belief that p but they formed the false second-order belief "I believe that not-p" or "I don't believe that p".

Unfortunately, in spite of its virtues, epistemic accounts of self-deception face the following problem. Since epistemic accounts understand lack of self-knowledge as the result of an epistemic failure in the first-person process of belief-formation for which the subject is epistemically responsible, a dilemma in regard to the following desiderata of self-deception arises (desiderata accepted by Fernández himself[44]): the subject's *epistemic responsibility* (fourth desideratum) and the *persistence* of self-deception (third desideratum). The dilemma goes as follows. On the one hand, if epistemic accounts go *internalist about justification* (as Fernández partially does[45]) claiming that subjects can always have access to the grounds that

[44] See Fernández, 2013, pp. 191-192 for persistence, and 2012, p. 382; 2013, pp. 182-188 for responsibility or normativity.

[45] Fernández (2013, pp. 44-45) considers his account partially internalist and partially externalist with respect to justification. However, this qualification doesn't affect the dilemma presented here.

justify their beliefs, they explain epistemic responsibility, but they fail to explain the persistence of self-deception. On the other hand, if epistemic accounts go *externalist about justification* claiming that subjects need not have access to the grounds that justify their beliefs, they explain the persistence of self-deception, but they fail to explain epistemic responsibility.

On the one hand, let's suppose that epistemic accounts go *internalist* so that John can always have access to the grounds that justify his beliefs. Assuming that John has what he takes as appropriate grounds to believe that he is ill (i.e., the spot on the back of his shoulder), he is supposed to have appropriate grounds to believe that he *believes* that he is ill (according to the bypass procedure). However, due to his epistemic failure, John is supposed to end up forming the ungrounded second-order belief that he believes that he is healthy. Imagine that someone points out to John both the grounds that justify (according to his own criteria) the first-order belief that he is ill (i.e.: the spot on the back of his shoulder) and the conflicting behaviour that he manifests regarding medical issues (e.g., avoiding doctors because of anxiety, enhancing the policy of his health insurance, etc). The evidence pointed out to John shows that either he shouldn't believe that he is ill or he shouldn't believe that he believes that he is healthy. Then, why can't John revise his ungrounded second-order belief (i.e., the second-order belief that he believes that he is healthy) and overcome self-deception rapidly and easily? Under the internalist horn of the dilemma, John has access to what he takes as appropriate grounds to believe that he is ill and he has been able to use those grounds appropriately to form the first-order belief that he is ill. Thus, it seems that he should also be able to use those grounds appropriately to revise the ungrounded second-order belief that he believes that he is healthy and to form the grounded second-order belief that he believes that he is ill; especially when someone just pointed out to him with good evidence the irrationality or inconsistency of his first-order and second-order beliefs. Therefore, it seems that if epistemic accounts go *internalist* about justification, they explain epistemic responsibility (fourth desideratum) quite well, for the subject would be the author of an epistemic error in forming a belief whose grounds are accessible to him, but they have trouble explaining the persistence of self-deception against being corrected and overcome by self-deceivers.

On the other hand, let's suppose that epistemic accounts go *externalist* about justification by claiming that subjects *need not have access* to the grounds that justify their beliefs. Then, by hypothesis, there must be cases in which self-deceivers cannot have access to the grounds for their beliefs. This horn of the dilemma is aimed at those cases. In cases in which self-deceivers cannot have access to the grounds for their beliefs, it is not difficult to explain the persistence of self-deception. Since John cannot have access from the first-person deliberative perspective to the grounds for his beliefs, he cannot correct the ungrounded second-order belief that he believes that he is healthy when the epistemic error is pointed out to him because he doesn't have agential control over that belief. The process of belief-formation works by itself here, with the unfortunate result that it elicits the ungrounded second-order belief that gives rise to self-deception. However, it is difficult to see how epistemic responsibility could be explained in these cases. Doxastic agency, together with reason-responsiveness (i.e., the existence of a threshold of evidence to change one's mind), is a necessary condition of epistemic responsibility (McHugh, 2013). If the subject is not the agent of her beliefs and self-deception is the result of an epistemic failure in a belief-forming process that works independently of the subject's deliberation and assessment of the evidence, then the subject cannot be held epistemically responsible for being self-deceived. Thus, if the epistemic account goes externalist, it explains the persistence of self-deception, but not the epistemic responsibility of the subject in self-deception.

Therefore, epistemic accounts of self-deception can explain either the third desideratum of self-deception (i.e., persistence against criticism on the basis of truthful evidence) or the fourth desideratum of self-deception (i.e., the epistemic responsibility of the subject), but they cannot explain *both* at the same time.

3.3 The psychological-state accounts of self-deception

Psychological-state accounts of self-deception consider that self-deception consists in a *sui generis* psychological state (i.e., a mental state or a cluster of mental states) in which self-deceivers are once they become self-deceived (instead of in the process that generates that psychological state). Firstly, it has been argued (Bach, 1981) that self-deception consists in *avoidance of the distressing recurrent thought that not-p* (e.g., "I am ill") when the subject believes that not-p (e.g., that he is ill) but desires that p (e.g., to be healthy). Secondly, it has been argued (Audi, 1982, 1988, 1989, 1997; Rey, 1988) that self-deception consists in the *sincere avowal* that p (e.g., "I believe that I am healthy") without the corresponding belief that p (e.g., without the corresponding belief that one is healthy). However, these psychological-state accounts of self-deception fail to explain the conflicting behaviour characteristic of self-deceivers (first desideratum). Indeed, the conflicting behaviour characteristic of self-deception is between what the subject sincerely says (e.g., "I believe that I am healthy") and what the subject does (i.e., enhancing his health insurance, avoiding doctors' appointments, etc.), but *not only* between what the subject sincerely says and does. There are also non-verbal actions that are coherent with what the subject sincerely says (after all, the subject is *sincere* in spite of the conflicting behaviour) and incoherent with other actions of the subjects (i.e., the actions that conflicts with what the subject sincerely says). For instance, John is having a child because he and his partner have decided to, and John is very happy about it. Why shouldn't he be? He sincerely says that he is healthy. Neither of these accounts of self-deception, though, can explain why John has decided to have a child expecting to educate and raise him. On the one hand, if self-deception were *sincere avowal* without belief (Audi, Rey), John wouldn't have decided to have a child with the intention to raise and educate him. For sincere avowals are supposed to be limited to verbal behaviour (e.g., "I believe that I am healthy"), and so, they are not supposed to be responsible for non-verbal actions (e.g., having a child with the intention to raise and educate him). Otherwise, sincere avowals would be similar to beliefs, and the introduction of the idiom "sincere avowal" would not be justified. On the other hand, if self-deception were avoidance of the distressing thought that p while believing that p (Bach), maybe John would say "I believe that I am healthy" to avoid the recurrent thought that he is not healthy, but he certainly wouldn't have decided to have a child with the intention to educate and raise him.

The most recent psychological-state account of self-deception is Gendler's account of self-deception as pretence (2010, Ch. 8). According to the pretence account, a subject is self-deceived when she believes that p but *pretends* that not-p because she desires that not-p. Thus, John is supposed to be self-deceived because he believes that he is ill but pretends that he is healthy (avoiding so the anxiety-producing thought that he is ill) because he desires to be healthy. At first sight, this account of self-deception fails to explain the irrational conflicting behaviour characteristic of self-deceivers (first desideratum) in the same way that the other psychological-state accounts: if John were pretending that he is healthy while believing that he is ill, he wouldn't have decided to have a child expecting to educate and raise him. However, Gendler defends her account of self-deception on the basis of the idea that John-like cases are instances of *self-delusion* rather than instances of *self-deception*, and so, on the basis of the idea that John-like cases are not under the scope of her account of self-deception as pretence.

According to Gendler, the difference between cases of self-deception and cases of self-delusion is the following. While self-deceivers take precautions to avoid evidence that could override their deception because they are *pretending that p* without believing that p is the case, self-deluders don't take such precautions because they *actually believe that p*, and so, they have nothing to fear. To illustrate this difference, Gendler describes an example very similar to John's case. Imagine a subject who is obviously ill and who has some kind of irrational attitude about her disease (i.e., either self-deception or self-delusion). Imagine that a doctor offers to this person the option of taking a medicine that is known for healing people with the disease that she has and for causing dreadful collateral effects to people who don't have the disease that she has. According to Gendler, two things can happen from now on. On the one hand, if this person decides to take the medicine, she was *self-deceived* about her disease because she believed that she is ill all along while pretending that she was healthy. Since to take the medicine and get healed is a better way to satisfy her desire to be healthy than pretending to be healthy, she stops pretending and takes the medicine. On the other hand, if this person decides not to take the medicine, she is *self-deluded* about her disease because she genuinely believes that she is healthy in the face of obvious evidence to the contrary. Hence, from Gendler's interpretation of this example seems to follow that John is self-deluded (i.e., he believes that he is healthy) rather than self-deceived (i.e., pretending that he is healthy) if he decides to have a child expecting to raise and educate him, and so, it seems to follow that John's case is not under the scope of Gendler's account of self-deception.

However, two problems arise against Gendler's interpretation of John-like examples. Firstly, John-like examples have at the same time some features that Gendler considers characteristic of self-delusion and some features that Gendler considers characteristic of self-deception. On the one hand, John is confident enough about the fact that he is healthy to decide to have a child expecting to educate and raise him, and so, in Gendler's interpretation, he counts as self-deluded in that regard (for, if John were the subject of Gendler's example, he could have refused to take the medicine just as much as he decided to have a child). On the other hand, John also "takes precautious" to avoid evidence that could override the judgement that he is healthy because he is motivationally biased to judge that he is healthy (e.g., he avoids going to the doctor or looking at the spot in the mirror), and so, in Gendler's interpretation, he counts as self-deceived in that regard. Therefore, Gendler's characterization of self-deceivers and self-deluders seems to have missed the point: there are cases (i.e., John-like examples) which have features that Gendler considers characteristic of self-deceivers and features that Gender considers characteristic of self-deluders at the same time.

Secondly, Gendler's characterization of self-deluders as believers has been questioned in the literature about delusion so that the possibility of a different interpretation of self-delusion seems to remain open. Indeed, 1) it has been argued that delusions have to do with a *sui generis* mental state different from belief (Egan, 2008; Tumulty, 2012); 2) it has been suggested (Szabados, 1985, p. (160 DSM-V, 2000, p. 819) that subjects are deluded when they claim something in the face of "incontrovertible and obvious proof or evidence to the contrary" (DSM-V, 2000, p. 819) so that delusions wouldn't be reason-responsive states (a condition of epistemic responsibility); and 3) it has been pointed out (Silva et al., 1998; Soyka, 1995; Soyka & Schmidt, 2011) that there is empirical evidence suggesting that delusions have to do with psychiatric disorders or neurophysiological problems (e.g., schizophrenia, brain injuries, etc.) because they don't usually appear in subjects with normal cognitive capacities. As a result, the difference between self-deception and self-delusion seems better characterized as follows: 1) self-deception and self-delusion are two different kinds of psychological states (i.e., mental states or clusters of mental states) and both are different from belief; 2) self-deceived states are reason-responsive (a condition of epistemic responsibility), meaning that there is a threshold of evidence that self-deceivers can reach to overcome self-deception, while self-delusions are not reason-responsive, meaning that there is no such threshold of evidence; and 3) self-deception takes place in subjects with normal cognitive capacities, while self-delusions are indicative of some kind of pathology involving impaired cognitive capacities.

Then, there are reasons to think that the example proposed by Gendler would be better understood in a different way. Particularly, there are reasons to think that the key to determine whether the person of Gendler's example is self-deceived or self-deluded has nothing to do with whether she takes the medicine or not; to the contrary, it has to do with whether her irrational state is reason-responsive or not. Indeed, two situations are possible: that the irrational mental state is reason-responsive or that the irrational mental state is not reason-responsive. On the one hand, if it is true of the person of Gendler's example that there is a threshold of evidence that she could reach in order to overcome her irrational state (e.g., that the doctor gave her a thorough diagnosis of her disease so that she dropped her irrational state), her irrational state is reason-responsive and the subject is epistemically responsible for having that irrational state. Under this development of Gendler's example, the subject is self-deceived about her health condition even if she decides not to take the medicine while maintaining that she is healthy (assuming that all-things-considered she manifests the irrational conflicting behaviour characteristic of self-deception in other areas of her life). On the other hand, if it is true of the person of Gendler's example that there isn't any threshold of evidence that she could reach in order to overcome her irrational state, her irrational state is not reason-responsive and the subject is not epistemically responsible for having her irrational state. Under this development of Gendler's example, the subject is self-deluded about her health condition even if she decides to take the medicine while maintaining that she is healthy.

Therefore, if Gendler's characterization of self-deception and self-delusion is mistaken and John's example is a case of self-deception rather than a case of self-delusion, Gendler's account of self-deception as pretence fails to explain John's irrational conflicting behaviour (first desideratum): if John were pretending to be healthy while believing that he is ill, he wouldn't have decided to have a child with the intention of raising and educating him. As a result, Gendler's account of self-deception is at an explanatory disadvantage against other accounts of self-deception that manage to explain John's paradigmatic example in an appropriate way.

In the remainder of this chapter, the behavioural-expressivist account of self-deception that follows from the behavioural-expressivist account of Transparency is going to be developed. In particular, this chapter is going to propose a psychological-state account of self-deception according to which self-deception is a *sui generis* kind of mental state that belongs to the class of unconscious mental states. Afterwards, it is going to be argued that the behavioural-expressivist account of self-deception is the only account that meets all the desiderata of a good account of self-deception. And, finally, it is going to be shown that the behavioural-expressivist account of self-deception has the additional advantage of distinguishing some phenomena that are considered to be different but closely related to each other: wishful thinking, self-deception and delusion.

3.4 The relational and the non-relational concepts of truth

As was said before, the behavioural-expressivist account considers that mental states are *patterns* of expressive behaviour. These patterns of expressive behaviour are identical to a set of *episodes of expression* distributed in a certain way over a certain period of time. Each one of these episodes of expression has a certain *expressive content*. For instance, a smile has the expressive content of happiness when it is an episode of the expressive pattern of happiness; the utterance "Ow! My leg!" has the expressive content of pain when it is an episode of the expressive pattern of pain; or the utterance "It is raining" has the expressive content of the belief that it is raining when it is an episode of the expressive pattern of the belief that it is raining. The expressive content of every episode of mental state can be considered from two different perspectives: it can be considered as *presenting* an aspect of a mental state of the subject or it can be considered as being actually or possibly *related* to an item of the world (i.e., as having intentionality). Let's see some examples. A smile of happiness can be seen either as presenting an aspect of the mental state of happiness of the subject or as being directed to a friend who has just arrived at the party (i.e., as having intentionality). The utterance "Ow! My leg!" can be seen either as presenting an aspect of the mental state of pain of the subject or as being directed to the leg of the subject (i.e., as having intentionality). And the utterance "It is raining" can be seen either as presenting an aspect of the belief that it is raining of the subject or as asserting the fact that it is raining (i.e., as having intentionality).

The intentionality of an episode of expression can be of two kinds: *propositional intentionality* or *non-propositional intentionality*. On the one hand, when the intentionality of the episode of expression is non-propositional, the item of the world to which it is actually or possibly related is an object or "entity". For instance, the smile of happiness or the utterance of pain "Ow! My leg!" has a non-propositional kind of intentionality because the item of the world to which they are actually or possibly related is an object or entity of the world: the friend at whom the smile is directed and the leg which is sore, respectively. On the other hand, when the intentionality of the episode of expression is propositional, the item of the world to which it is actually or possibly related is a fact or "state of affairs". For instance, the assertion "It is raining" has a propositional kind of intentionality because the item of the world to which it is actually or possibly related is a fact or state of affairs rather than an object or entity: the fact that it is raining.

Due to the fact that the expressive content of an episode of mental state can be seen both as *presenting* an aspect of the mental state of the subject and as being actually or possibly *related* to an item of the world (i.e., as having intentionality), there are two different senses in which the expressive content of an episode of mental state can be said to be true or false: the *relational sense* of truth and the *non-relational sense* of truth. On the one hand, the relational sense of truth (henceforth: *truth*$_{rs}$) is predicated of the expressive content of an episode of mental state insofar as it has intentionality; i.e., insofar as it is actually or possibly related with a certain object/entity or fact/state-of-affairs of the world. When the appropriate relation of fit between the expressive content of the episode of mental state and the object or fact of the world that it is about takes place, the expressive content is true in the relational sense (henceforth: *true*$_{rs}$). By contrast, when the appropriate relation of fit between the expressive content of the episode of mental state and the object or fact of the world that it is about doesn't take place, the expressive content is false in the relational sense (henceforth: *false*$_{rs}$). Let's continue with the examples. A smile of happiness at a friend who just arrived at the party is true$_{rs}$ when the person that the subject is seeing is actually the friend at whom she thinks that she is smiling (i.e., the appropriate relation of fit between the expressive content and its intentional object takes place) and it is false$_{rs}$ when the subject is mistaken and she is seeing a different person who happens to have the same haircut as the friend at whom she thinks she is smiling (i.e., the appropriate relation of fit between the expressive content and its intentional object doesn't take place). The utterance "Ow! My leg!" is true$_{rs}$ when the subject actually has the leg that is supposed to be sore (i.e., the appropriate relation of fit between the expressive content and its intentional object takes place) and it is false$_{rs}$ when the subject doesn't have the leg that it is supposed to be sore (this happens in cases of *phantom pain*: the subject feels pain in a limp that he doesn't have anymore because it was amputated). And the utterance "It is raining" is true$_{rs}$ when it is a fact that it is raining in the area that the utterance refers to (i.e., the appropriate relation of fit between the expressive content and its intentional fact takes place) and it is false$_{rs}$ when it is not a fact that it is raining in the area that the utterance refers to (i.e., the appropriate relation of fit between the expressive content and its intentional fact doesn't take place).

Notice that the concept of knowledge in the epistemic sense (*knowing that*) involves a warranted *belief* that is true in the relational sense (i.e., *true*$_{rs}$). Since beliefs have a propositional kind of intentionality, expressive episodes with a propositional kind of intentionality (i.e., assertions), rather than expressive episodes with a non-propositional kind of intentionality (e.g., smiles or expressions of pain like "Ow! My leg!"), are supposed to be possible instances of epistemic knowledge. Then, true$_{rs}$ assertions are the episodes of expression which are supposed to be involved in cases of knowledge of the world (e.g., "It is raining"), in cases of knowledge of other people's mental states (e.g., "Lydia believes that it is raining") and in cases of third-person epistemic self-knowledge (e.g., Tom's third-person avowal "I don't believe in gender equality").

On the other hand, the non-relational sense of truth (henceforth: *truth*$_{nrs}$) is predicated of the expressive content of an episode of mental state insofar as it is a *presentation* of a mental state of the subject. The expressive content of an episode of mental state is non-relationally true (henceforth: *true*$_{nrs}$) when the episode of expression is presented or shown in a *clear way* (i.e., as being *also in appearance* an episode of the mental state that it is actually an episode of). By contrast, the expressive content of an episode of mental state is non-relationally false (henceforth: *false*$_{nrs}$) when the episode of expression is presented or shown in a *misleading way* (i.e., as being *in appearance* an episode of a different mental state from the mental state that it is actually an episode of). Since truth$_{nrs}$ is the concept of truth that applies to episodes of expression insofar as they are either clear or misleading presentations of the subject's mental states, it is a concept of truth characteristic of the first-person perspective that has nothing to do with knowledge in the sense of *knowing that*. There are (at least) three reasons why the expressive content of an episode of mental state can end up being false$_{nrs}$: 1) the subject might be untruthful or insincere, 2) the subject might express a mental state with a wrong vehicle of expression, and 3) the expressed mental state might be an unconscious mental state (not to be confused with expressing a mental state un-self-consciously). Let's tackle them in order.

Firstly, 1) being truthful or untruthful, sincere or insincere, has to do with whether the subject *intentionally* tries to produce a clear (i.e., truth$_{nrs}$) or a misleading (i.e., false$_{nrs}$) episode of expression. When the subject manifests an episode of mental state with the intention to present or show its expressive content in a clear way (i.e., with the intention to present or show the expressive episode as being *also in appearance* an episode of the mental state that it really is an episode of), the expressive content of that episode of mental state is *truthful* or *sincere* (i.e., among its expressive content is the intention of the subject to be clear). If the intention of the subject is fulfilled and she manages to manifest a clear episode of expression, the expressive content of that episode will be true$_{nrs}$. For instance, my utterance "It is raining" is truthful or sincere when I produce it with the intention to be clear, that is, when I produce it with the intention to present or show its expressive content in a clear way: as being *also in appearance* an episode of my belief that it is raining. If I actually manage to be clear, the expressive content of my utterance will be true$_{nrs}$. By contrast, when the subject manifests an episode of mental state with the intention to present or show its expressive content in a misleading way (i.e., with the intention to present or show the expressive episode as being *in appearance* an episode of a *different* mental state from the mental state that it really is an episode of), the expressive content of that episode of expression is *untruthful* or *insincere* (i.e., among its expressive content is the intention of the subject to be misleading). If the intention of the subject is fulfilled and she manages to produce a misleading episode of expression, its expressive content will be false$_{\text{nrs}}$. For instance, my utterance "It is raining" is untruthful or insincere when I manifest it with the intention to be misleading, that is, when I produce it with the intention to present or show its expressive content in a misleading way: as an *apparent* episode of the belief that it is raining when it is actually an episode of the mental state of pretending to believe that it is raining. If I actually manage to produce a misleading episode of expression, its expressive content will be false$_{\text{nrs}}$.

The following two examples of *intentional pretence* should clarify the notion of *truthfulness* or *sincerity* further. Imagine that I moan apparently out of pain because I am pretending that I have a headache in front of you to rid myself from cleaning the house. Since I am pretending to have a headache to deceive you, I will intentionally try to produce a moan as similar in appearance as possible (i.e., in tone, in duration, in facial expression, in bodily posture…) to a moan of a real headache so that I am successful at deceiving you. Then, I am being untruthful and insincere because I intentionally try to present my moan as an *apparent* expressive episode of headache when it is actually an expressive episode of *pretended* headache. If I manage to produce a misleading moan in this sense (i.e., an *apparent* moan of headache), its expressive content will be false$_{\text{nrs}}$. By contrast, imagine that I moan in front of you to make the *joke* that I am pretending to have a headache in order to rid myself from cleaning the house. Since I am pretending to have a headache to make a joke and not to deceive you, I will try to produce a moan which is somehow *different in appearance* (i.e., in tone, in duration, in facial expression, in bodily posture…) from a real moan of headache so that you know that I am joking. Then, I am being truthful and sincere because I intentionally try to present my moan as a clear episode of what it really is: an expressive episode of a joke instead of an expressive episode of an actual headache. If I manage to satisfy my intention, the expressive content of my moan will be true$_{\text{nrs}}$.

Therefore, the expressive content of an episode of mental state can be truthful (sincere) or untruthful (insincere) depending on how its expressive content is *intended* to be presented or shown by the subject: either in a clear or in a misleading way. As a result, the expressive content of an episode of mental state can be true$_{\text{nrs}}$ or false$_{\text{nrs}}$ depending on whether the subject manages to satisfy her intention or not, and so, depending on whether she manages to be clear or misleading. Notice that expressing a false$_{\text{nrs}}$ episode of mental state as a result of being untruthful or insincere doesn't involve any lack of first-person self-knowledge (*knowing how*).

On the one hand, being untruthful or insincere doesn't involve lack of first-person expressive self-knowledge (*knowing how*) in the sense of not having the ability to express a mental state because being untruthful or insincere requires having the ability to express the mental state of *pretending M* in an appropriate way (i.e., as if one were expressing the mental state M itself). On the other hand, being untruthful or insincere doesn't involve lack of first-person expressive self-knowledge (*knowing how*) in the sense of expressing that mental state un-self-consciously because it is clear that one can (and most of the time will) exercise self-consciously (i.e., attentively and knowing what one is up to) the ability to *pretend* the mental state M; i.e., the ability to produce expressive episodes as similar in appearance as possible to the mental state M itself.

Secondly, 2) the expressive content of an episode of mental state can be presented or shown in a right vehicle of expression (i.e., as it is characteristic of the expressive pattern of the mental state in the given community) or in a wrong vehicle of expression (i.e., as it is not characteristic of the expressive pattern of the mental state in the given community), regardless of the intention of the subject. When the expressive content of an episode of mental state is presented or shown in an appropriate vehicle of expression (i.e., as it is required by the pattern of expression of the mental state), the episode of mental state is an *expressive success* in that regard and it could be true$_{\text{nrs}}$ because its expressive content could be presented or shown in a clear way. By contrast, if the expressive content of an episode of mental state is presented or shown in an inappropriate vehicle of expression (i.e., as it is not allowed by the pattern of expression of the mental state), the episode of mental state is an *expressive failure* in that regard and it will be false$_{\text{nrs}}$ because its expressive content will be presented or shown in a misleading way.

Consider Bar-on's example of the dentist, which is a case of expressive failure due to fear. I am in the dentist's chair waiting for the dentist to drill my tooth. I open my mouth, the dentist brings the drill closer to my tooth and I sincerely shout "I feel a terrible pain!" or "Stop! It is painful!" before the drill actually touches my tooth. Since the dentist didn't reach my tooth, I am not in pain. So, my shout is not an expressive episode of pain but an expressive episode of fear (for it is fear and not pain what I have). In this case, the shout "I feel a terrible pain!" or "Stop! It is painful!" presents or shows its expressive content in an inappropriate way: "I feel a terrible pain!" or "Stop! It is painful!" are *vehicles of expression* characteristic of the expressive pattern of pain and not of the expressive pattern of fear. Then, the shouts "I feel a terrible pain!" or "Stop! It is painful!" are false$_{\text{nrs}}$ because they present or show their expressive content in a misleading way: their expressive content is fear because they are episodes of fear, but their expressive content is presented or shown in vehicles of expression characteristic of the expressive pattern of pain and not of the expressive pattern of fear. Compare this example with a case in which the drill actually touches my tooth and the anaesthesia didn't take effect fast enough, so I shout "I feel a terrible pain!" or "Stop! It is painful!". In this case, the shouts present or show their expressive content in an appropriate way: their expressive content is pain because they are episodes of pain and their expressive content is presented or shown in vehicles of expression characteristic of the expressive pattern of pain. Then, the shouts "I feel a terrible pain!" or "Stop! It is painful!" are true$_{nrs}$ because they present or show their expressive content in a clear way.

Notice that producing a false$_{nrs}$ expressive episode as a result of using the wrong vehicle of expression is compatible with having first-person expressive self-knowledge in the sense of having the *ability* to express that mental state, but it involves lack of first-person expressive self-knowledge in the sense that the false$_{nrs}$ expressive episode is necessarily *un-self-consciously* expressed. On the one hand, making the mistake of using a wrong vehicle of expression is compatible with having first-person expressive self-knowledge in the sense of having the ability to express that mental state because one can make a mistake when expressing a mental state (especially if the circumstances are abnormal, as in the dentist's example) without lacking the ability to appropriately express that mental state (for some mistakes are normal even in subjects who master the ability to appropriately express a mental state). On the other hand, making the mistake of using a wrong vehicle of expression involves lack of first-person expressive self-knowledge in the self-consciousness sense because one cannot express self-consciously (i.e., knowing what one is up to) a mental state with a wrong vehicle of expression. Continuing with the example, I might have tried to attentively exercise my ability to express in an appropriate way any sensation of pain that I could have had during the dental appointment (making it so the dentist stops as soon as possible if I felt pain), but when I said "I feel a terrible pain!" or "Stop! It is painful!" before the dentist even touches my tooth, I didn't know what I was up to (*knowing how*), for I expressed my fear with a vehicle of expression of pain rather than with a vehicle of expression of fear. As a result, I have first-person expressive self-knowledge in the sense that I still qualify as having the ability to express my mental state of pain in an appropriate way (even if I made a mistake on the given occasion), but I lack first-person expressive self-knowledge in the self-conscious sense because "I feel a terrible pain!" or "Stop! It is painful!" are un-self-conscious episodes of expression (i.e., I didn't know what I was up to when I produced those expressive episodes).

Therefore, to manifest a false$_{\text{nrs}}$ avowal in the type of cases described in 1) and 2) doesn't necessarily involve having a complete lack of first-person expressive self-knowledge in the *knowing how* sense. 1) In the case of false$_{\text{nrs}}$ episodes of expression resulting from the insincerity or untruthfulness of the subject, no lack of first-person self-knowledge is necessarily involved at all. And 2) in the case of false$_{\text{nrs}}$ episodes of expression resulting from the mistake of expressing a mental state with the wrong vehicle of expression (expressive failure), the subject lacks first-person self-knowledge in the sense that the false$_{\text{nrs}}$ expressive episode is un-self-consciously expressed (i.e., without knowing what one is doing) but not necessarily in the sense that she lacks the ability to express that mental state (for making some mistakes is compatible with having the ability to appropriately express a mental state). Things are different, though, in the case of unconscious mental states, where a complete lack of first-person expressive self-knowledge is necessarily involved.

The distinction between conscious and unconscious mental states has to do with the distinction between implicit and explicit linguistic expression. Linguistic expressions can be *implicit* or *explicit*[46]. They are implicit when they *don't mention* the mental state that they are allegedly episodes of (e.g., "It is going to rain", "If you came back soon…" or "I'd eat a whole cow!"). And they are explicit when they *mention* the mental state that they are allegedly episodes of (e.g., avowals like "I *believe* that it is going to rain", "I *wish* you came back soon" or "I *feel* hungry"). In turn, explicit linguistic expressions (i.e., avowals) can mention either the mental state that they are actually episodes of or a mental state different from the mental state that they are actually episodes of. When they mention the mental state that they are actually episodes of, they (normally) clarify further their expressive content, and so, they are true$_{\text{nrs}}$ (e.g., saying "I *believe* that it is raining" to express my belief that it is raining). By contrast, when they mention a mental state different from the mental state that they are actually episodes of, they (normally) blur their expressive content, and so, they are (normally)[47] false$_{\text{nrs}}$ (e.g.,

[46] The use of "explicit" and "implicit" made here should be understood analogously with the use made by Austin (1962, 1970) in regard to speech acts. According to Austin, an explicit utterance (e.g., "I *order* you to clean the house") is an utterance in which the illocutionary force (assertion, order, warning…) of the utterance is clarified by being explicitly mentioned. Likewise, it will be considered here that an explicit linguistic expression (e.g., I *believe* that it is raining) is an utterance in which the mental state that it is an expressive episode of is clarified by being explicitly mentioned.

[47] For remember the example in which I make the joke that I am pretending to have a headache in order not to clean the house. In this case, I can say "I have a terrible headache" (making so explicit the wrong mental state

saying "I *believe* that it is raining" to express my mental state of pretending that I believe that it is raining).

So, thirdly 3), conscious mental states are those mental states that can be expressed with first-person utterances in present tense that make explicit the *right* mental state (i.e., that makes explicit the mental state that they are *actually* an expressive episode of), while unconscious mental states are those mental states that cannot be expressed with first-person utterances in present tense that make explicit the *right* mental state (i.e., that make explicit the mental state that they are *actually* an expressive episode of). In other words, conscious mental states are those mental states that can be expressed with true$_{nrs}$ avowals (i.e., with avowals that make explicit the mental state that they are actually expressive episodes of), while unconscious mental states are those mental states that *cannot* be expressed with true$_{nrs}$ avowals (i.e., with avowals the make explicit the mental state that they are actually expressive episodes of). On the one hand, the reason why conscious mental states can be expressed with true$_{nrs}$ avowals is that true$_{nrs}$ avowals are expressive episodes that belong to the characteristic expressive patterns of conscious mental states. Belief, desire, intention, pain, pretence… are examples of conscious mental states because their true$_{nrs}$ avowals belong to their characteristic pattern of expression (e.g., "I believe that p", "I desire that p", "I intend that p", "I have a pain in my leg" and "I am pretending that p", respectively). All of those mental states are conscious mental states because they can be expressed with a true$_{nrs}$ avowal; i.e., with an avowal the makes explicit the mental state that they are actually an expressive episode of (e.g., belief, desire, intention, pain and pretence). On the other hand, the reason why unconscious mental states cannot be expressed with true$_{nrs}$ avowals is that true$_{nrs}$ avowals are not expressive episodes that belong to the characteristic expressive patterns of unconscious mental states. As a result, unconscious mental states can only be expressed with false$_{nrs}$ avowals because only false$_{nrs}$ avowals (i.e., avowals that makes explicit a mental state different from the mental state that they are actually episodes of) belong to their characteristic patterns of expression. As it will be argued in the following sections, self-deception and delusions are examples of unconscious mental states, for the true$_{nrs}$ avowal "I am self-deceived about p" or "I am deluded about p" are not expressive episodes that belong to the expressive pattern of self-deception or to the expressive pattern of delusion. Thus, self-deception and delusion can only be expressed with false$_{nrs}$ avowals (e.g., "I believe that p").

because I don't have a headache) without my explicit linguistic expression being false$_{nrs}$ because of that (the context leaves clear that the expression is an episode of a joke).

As a result, conscious and unconscious mental states are two different *classes* or *genera* of mental states because they have two different *classes* or *genera* of expressive patterns. Conscious mental states have true$_{\text{nrs}}$ avowals among the expressive episodes of their patterns of expression, and unconscious mental states only have false$_{\text{nrs}}$ avowals among the expressive episodes of their patterns of expression. Then, the difference between conscious and unconscious mental states is *ontological* rather than *epistemic*: it is a matter of belonging to the class of conscious mental states or to the class of unconscious mental states, and not a matter of lacking epistemic self-knowledge (i.e., true warranted belief) of a particular mental state. So, the process of overcoming an unconscious mental state should be understood here as the process of *replacing* a mental state that belongs to the class of unconscious mental state by a different mental state that belongs to the class of conscious mental states rather than as the process of acquiring epistemic self-knowledge of an unconscious mental state.

Since conscious mental states can be expressed with true$_{\text{nrs}}$ avowals, conscious mental states can be self-consciously expressed (i.e., knowing what one is up to). However, since unconscious mental states cannot be expressed with true$_{\text{nrs}}$ avowals, unconscious mental states can only be expressed un-self-consciously (i.e., without knowing what one is up to). On the one hand, conscious mental states can be expressed self-consciously because, thanks to the fact that they can be expressed with true$_{\text{nrs}}$ avowals, subjects can exercise the activities that typically express a conscious mental state (e.g., making an assertion, picking up the umbrella, eating an ice cream) knowing what they are up to or knowing what they are doing; i.e., knowing the kind of activity that they are performing and the mental state that they are expressing. Of course, a conscious mental state might be un-self-consciously expressed sometimes (e.g., distractedly and without knowing what one is up to), but the characteristic feature of conscious mental states is that they *can* be self-consciously expressed. For instance, my belief that a certain politician has bad ethical standards is a conscious mental state because, insofar as its true$_{\text{nrs}}$ avowal (e.g., "I believe that…) belongs to its characteristic pattern of expression, it can be self-consciously expressed. Indeed, even if I might sometimes express my belief without knowing what I am doing (e.g., with a subtle grimace of disgust that I unwillingly make when hearing someone talking about him), I can express my belief knowing the kind of activity that I am performing and the mental state that I am expressing as well. For instance, asserting "I believe that such politician has bad ethical standards" and proceeding to explain step by step the reasons that support my belief.

On the other hand, unconscious mental states cannot be expressed self-consciously because, insofar as they can only be expressed with false$_{\text{nrs}}$ avowals, they cannot be expressed knowing what one is up to. Indeed, due to the fact that true$_{\text{nrs}}$ avowals don't belong to the expressive pattern of unconscious mental states, subjects cannot exercise the activity that expresses an unconscious mental state while knowing what they are doing; i.e., while knowing the kind of activity that they are performing and the mental state that they are expressing. As a result, even when subjects *attentively* perform an activity that involves the expression of an unconscious mental state, they don't *know what they are doing* because they exercise that activity as if they were performing a different one: as if they were performing an activity characteristic of the *conscious* mental state (e.g., belief, desire, intention, pretence, etc.) explicitly mentioned in the false$_{\text{nrs}}$ avowal that belongs to the expressive pattern of the unconscious mental state in question. As it will be argued in the next section, self-deception is an example of unconscious mental state, and so, it cannot be expressed self-consciously because the true$_{\text{nrs}}$ avowal "I am self-deceived about the fact that I am healthy" is not a possible expressive episode of self-deception. As a result, a self-deceived subject might *attentively* perform an activity that involves the expression of his unconscious mental state of self-deception (e.g., saying "I am healthy" or "I am too busy to go to the doctor", enhancing his health insurance, etc.) without *knowing what he is up to*. For he performs those activities as if he were expressing conscious mental states (i.e., the belief that he is healthy, the belief that he is too busy, the belief that he can afford his health insurance, etc.) when in fact he is expressing a mental state of self-deception.

Therefore, subjects can have first-person expressive self-knowledge (*knowing how*) of conscious mental states, but they cannot have first-person expressive self-knowledge of unconscious mental states. Subjects can have first-person self-knowledge of conscious mental states both in the sense that they can have the ability to express them in an appropriate way and in the sense that they can express them self-consciously (i.e., knowing what they are up to). However, subjects cannot have first-person self-knowledge of unconscious mental states in any of those two senses. Firstly, subjects can only express an unconscious mental state un-self-consciously. Secondly, insofar as unconscious mental states can only be expressed un-self-consciously, it is not possible that subjects can have the ability to express an unconscious mental state in an appropriate way: they are always inappropriately expressed because they are always expressed without knowing what one is up to or what one is doing.

In summary, the expressive content of an episode of mental state can be considered from two different perspectives. When the expressive content of an episode of mental state is considered from the perspective of the *relation* that it has to a certain object or fact of the world (i.e., as having intentionality), the expressive episode is either true$_{rs}$ or false$_{rs}$ depending on whether the appropriate relation of fit takes place or not. By contrast, when the expressive content of an episode of mental state is considered from the perspective of being a *presentation* of a mental state of the subject, the expressive episode is either true$_{nrs}$ or false$_{nrs}$ depending on whether it is presented or shown in a clear way (i.e., as having, also in appearance, the expressive content that it actually has) or in a misleading way (i.e., as having in appearance an expressive content different from the expressive content that it actually has). Among the elements that can affect the "truth$_{nrs}$-value" of an expressive episode are: 1) whether the subject is truthful (sincere) or untruthful (insincere), 2) whether the vehicle of expression used by the subject is appropriate (expressive success) or not (expressive failure) and 3) whether the expressed mental state is conscious or unconscious.

3.5 The behavioural-expressivist account of self-deception

In this section a new psychological-state account of self-deception is going to be developed on the basis of the behavioural-expressivist account of Transparency. Particularly, it is going to be argued that from the behavioural-expressivist account of Transparency follows that self-deception is a *sui generis* mental state that involves both a lack of first-person expressive self-knowledge (*knowing how*) and difficulties to acquire third-person epistemic self-knowledge (*knowing that*). Mental states of self-deception will be called *self-deceived mental states* henceforth.

To remember, the behavioural-expressivist account of Transparency claims that the question "Do you believe that p?" can be meant in a deliberative or in a self-ascriptive way. When the question "Do you believe that p?" is meant in a deliberative way, it is transparent to the question "Is p the case?" because it is a question about whether p. As a result, the subject is supposed to answer that question with the judgement that p is the case (i.e., "I believe that p" or "p is the case"), with the judgement that not-p is the case (i.e., "I believe that not-p" or "p is not the case") or with a suspension of judgement (i.e., "I don't believe either way" or "It could be either way") made at the end of a *first-person deliberative process* about whether p. By contrast, when the question "Do you believe that p?" is meant in a self-ascriptive way, it is not transparent to the question "Is p the case?" because it asks about whether the subject believes that p and not about whether p. As a result, the subject is supposed to answer that question with a self-ascription of belief (e.g., "I believe that p"), a self-ascription of disbelief (e.g., "I don't believe that p") or a self-ascription of lack of belief (e.g., "I don't believe either way) made at the end of a *third-person process of self-inspection* based on evidence about the subject's own mental states. In what follows, it is going to be argued that self-deception can be explained from the behavioural-expressivist account of Transparency attending to the differences between what happens when a self-deceived subject deliberates from the first-person perspective about the issue that he is getting self-deceived about (e.g., "Do you believe that you are healthy?" meant in a deliberative way) and what happens when a self-deceived subject self-inspects himself to find out which are his mental states about the issue that he is self-deceived about (e.g., Do you actually believe that you are healthy?" meant in a self-ascriptive way). Let's explain this idea using John's case as a paradigmatic example.

When John finds the spot on the back of his shoulder, he finds it disturbing and worrisome. Then, he asks himself the first-person deliberative question "Can this spot be malignant? Am I healthy?". To answer this question, John deliberates about whether the spot is malignant or not on the basis of evidence (e.g., its shape and colour). John desires to be healthy, and so, he finds disturbing the possibility of being ill and relieving the possibility of being healthy. Since he doesn't know for how long the spot has been there, the possibility of being ill makes him especially nervous because he is not sure about whether the spot would respond to treatment in the hypothetical case that it is malignant. Then, when collecting and assessing the evidence about whether the spot is malignant or not, he is motivationally biased to judge that he is healthy. For instance, he pays more attention and gives more weight to the colour of the spot (which he finds relieving because he finds it similar to the colour of a regular mole) than to the shape of the spot (which he finds distressing because he finds it less similar to the shape of a regular mole). As a result of this motivated bias, John hastily concludes his deliberation with the biased and relieving judgement "This is just a regular mole. So, [I believe that] I am healthy". Normally, this judgement would have given rise to the *conscious beliefs* that the spot is just a regular mole and that he is healthy. However, in this case, this judgement gives rise to the (unconscious) *self-deceived mental state* that the spot is just a regular mole and that he is healthy. For, due to the motivated bias, the following *expressive failure* occurs: John's judgement "This is just a regular mole. So, [I believe that] I am healthy" is an expressive episode of the false$_{nrs}$ avowal that gives rise to the (unconscious) self-deceived mental state that the spot is just a regular mole and that he is healthy (rather than being an expressive episode of the true$_{nrs}$ avowal that gives rise to the conscious belief that he is healthy). Once John has formed the self-deceived mental state that he is healthy, he will start to express the pattern characteristic of the self-deceived mental state that he is healthy (among which are further episodes of motivated bias). For instance, he will skip doctors' appointments and chats about medical issues, he will enhance his health insurance, he will get interested in religious topics, he will try to justify his irrational behaviour making up sincere excuses, and so on. Since the self-deceived mental state of John resembles belief (e.g., he sincerely says "I *believe* I am healthy"), it can be called *self-deceived belief*.

It is important to notice that the existence of a motivated bias in John's first-person deliberation about whether he is healthy is independent of whether John is healthy or not as a matter of fact. For the existence of a motivated bias in John's deliberation is independent of whether the evidence supports the fact that he is healthy or the fact that he is ill. Indeed, in order for a motivated bias to take place, it is enough that John collects and assesses the evidence in a way that goes against his own *epistemic norms* because of a motivational state (e.g., the desire to be healthy, anxiety about the possibility of being ill, etc.), regardless of whether John is healthy or not, and so, regardless of whether the evidence actually supports that John is healthy or not. Therefore, regardless of whether the spot is malignant or not, John is motivationally biased when he deliberates about the character of the spot because he collects and assesses the evidence in a way that goes against his own epistemic norms: he avoids paying attention to the shape of the spot out of anxiety, he gives more weight to the colour of the spot because he finds relieve in doing so, he avoids thinking that the spot might have to be checked by a doctor out of anxiety, etc.. And even after having formed the self-deceived belief that he is healthy, he keeps expressing episodes of motivated bias because, against his own epistemic norms, John avoids paying attention to the possible changes that the spot might have undertaken out of anxiety, avoids doctors' appointments and talks about medical issues out of anxiety, and so on.

To see why self-deception involves lack of self-knowledge both in the first-person expressive sense of *knowing how* and in the third-person epistemic sense of *knowing that* (i.e., true$_{rs}$ warranted belief), let's see how John could answer Lydia's questions. When Lydia asks

John “Do you believe that you are healthy?” after observing his conflicting behaviour, John can answer “[I believe that] I am healthy” either from the first-person deliberative perspective (in which Transparency takes place) or from the third-person self-inspective perspective (in which Transparency doesn’t take place). On the one hand, when John answers from the deliberative first-person perspective, he answers with a judgement about whether he is healthy on the basis of no extra reasons about his health condition (for he already deliberated and made up his mind about the issue before, when he saw the spot). Particularly, John answers with the judgement “[I believe that] I am healthy. I’m as fit as a fiddle!”. However, this judgement is an un-self-conscious episode of expression because John exercises the ability to make a judgement about the issue and to express his (unconscious) *self-deceived belief* as if he were exercising a different activity; i.e., the activity of expressing his (conscious) *belief* about his health condition. So, John doesn’t know what he is doing when he answers with the judgment “[I believe that] I am healthy. I’m as fit as a fiddle!”. In fact, since John’s self-deceived belief is an unconscious mental state (i.e., its true$_{nrs}$ avowal “I am self-deceived about the fact that I am healthy” is not an expressive episode belonging to its expressive pattern), it can only be expressed un-self-consciously (i.e., without knowing what one is up to).

Thus, each time when John performs an activity that involves the expression of his self-deceived belief, he *doesn’t know what he is up to* because he exercises that activity as if he were performing a different activity: as if he were performing an activity that involves the expression of *a conscious belief* rather than the expression of a *self-deceived belief.* When John attentively skips his doctor’s appointments saying “[I believe that] I’m too busy to go to the doctor”, he doesn’t know what he is up to because he is exercising an activity that expresses his *self-deceived belief* that he is healthy as if he were performing an activity that expresses the *belief* that he is too busy to go to the doctor. Or when John attentively enhances his health insurance saying “[I believe that] I can afford it”, he doesn’t know what he is up to because he is exercising an activity that expresses his *self-deceived belief* that he is healthy as if he were performing an activity that expresses the *belief* that he can afford his health insurance.

As a result, there isn’t any sense of first-person expressive self-knowledge (*knowing how*) in which John could have first-person expressive self-knowledge when he expresses his self-deceived belief that he is healthy. On the one hand, his self-deceived belief is an unconscious mental state that cannot be self-consciously expressed by John because subjects cannot perform the activity to express an unconscious mental state knowing what they are up to. On the other hand, since John’s self-deceived belief that he is healthy cannot be self-consciously expressed, it is not possible that he has the ability to express his self-deceived belief in an appropriate way: each time that John expresses his self-deceived belief, he expresses it without knowing what he is doing (i.e., un-self-consciously).

On the other hand, when John answers Lydia's question "Do you believe that you are healthy?" from the third-person self-inspective perspective, he answers with a self-ascription or judgement about his mental states made on the basis of evidence about his own attitudes. If John manages to make the true$_{rs}$ self-ascription of mental state "I am self-deceived about my health condition" (i.e., third-person avowal) or "It is a fact that I am self-deceived about my health condition" (i.e., assertion) and if he makes it on the basis of good evidence, he will form a true$_{rs}$ warranted second-order belief, acquiring so third-person epistemic self-knowledge of the fact that he has a self-deceived belief. However, managing to acquire third-person epistemic self-knowledge of a self-deceived mental state is more difficult than managing to acquire third-person epistemic self-knowledge of conscious mental states for two reasons. Firstly, some of the expressive episodes characteristic of self-deceived mental states *resemble in appearance* some of the expressive episodes characteristic of conscious attitudes. Secondly, the same motivational state (e.g., the desire to be healthy, anxiety about the possibility of being ill, etc.) that biases the first-person deliberation about whether p that gives rise to self-deception, biases the subject's self-inspective process about whether he has such a self-deceived mental state as well. Let's explain them in order.

On the one hand, some of the expressive episodes of self-deceived mental states are similar in appearance to the expressive episodes of some conscious attitudes. Particularly, since self-deceived mental states are unconscious mental states, all the *linguistic expressive episodes* of a self-deceived mental state are similar in appearance to the *linguistic expressive episodes* of other conscious attitudes (e.g., belief). For instance, some of the linguistic expressive episodes of John's self-deceived belief that he is healthy are similar in appearance to the linguistic expressive episodes of the belief that he is healthy (e.g., "[I believe that] I am healthy"); and others of the linguistic expressive episodes of John's self-deceived belief are similar in appearance to the linguistic expressive episodes of others beliefs: the belief that he is too busy to attend his medical appointment (e.g., "I am going to skip my medical appointment because I am too busy today"), the belief that it is important to be cautious in life or the belief that he can afford the health insurance (e.g., "I am going to enhance my health insurance because it is important in life to be cautious and I can afford it").

Then, when John self-inspects himself from the third-person perspective to find out which attitude he holds about his health condition on the basis of the evidence about his own mental states provided by the epistemic sources of memory and introspection, he finds the following pieces of evidence: that he sincerely said in the past "I believe that I am healthy" (memory), that he skipped his medical appointment because he sincerely thought "I am too busy today" (memory), that he enhanced his health insurance because he sincerely thought "It is important in life to be cautious and I can afford it" (memory), that he would think again "I am too busy to go to the doctor today" or "I can afford the insurance" in similar circumstances (introspection), and so on. As a result, since these expressive episodes of self-deception are similar in appearance to expressive episodes of other mental states (i.e., the belief that he is healthy, the belief that he was busy, the belief that it is important to be cautious and the belief that he can afford the health insurance), it is likely that John concludes the process of self-inspection by mistakenly judging "I believe that I am healthy" rather than "I have the self-deceived belief that I am healthy", forming so the false$_{rs}$ second-order belief that he believes that he is healthy and failing to obtain third-person epistemic self-knowledge.

On the other hand, since self-deceived mental states are caused by a motivational state, this motivational state (e.g., the desire to be healthy, anxiety about the possibility of being ill, etc.) may also bias the subject's third-person process of self-inspection about which is the attitude that he holds about p. For instance, when John self-inspects himself to find out the attitude that he holds about his health condition on the basis of the evidence about his mental states provided by memory or introspection, he is motivationally biased to collect and assess the evidence in a way that favours the judgement that he believes that he is healthy because, due to his desire to be healthy, he finds *relief* in thinking that he has the belief that he is healthy and *anxiety* in thinking that he might be self-deceived about the fact that he is healthy. For thinking that one is self-deceived about the fact that one is healthy is the first step of doubting whether one is healthy as a matter of fact.

This motivated bias in the third-person process of self-inspection goes as follows. On the one hand, John finds it *disturbing* to pay attention to the evidence that *apparently* counts in favour of the fact that he has the self-deceived belief that he is healthy. So, he tends to avoid paying attention to the fact that he mustn't have been that busy when he skipped his doctor's appointment if he was up for a beer with Lydia (memory) or to the fact that it must have been obvious to him that he wasn't doing that well financially if he had to borrow some money from Lydia to pay the rent (memory). On the other hand, John finds *relief* in paying attention to the evidence that *apparently* counts in favour of the fact that he believes that he is healthy. So, he tends to pay more attention to the fact that he has always said "I'm as fit as a fiddle" (memory), to the fact that he would think again "I am too busy to go to the doctor today" in similar circumstances (introspection), or to the fact that he is having a child and he is very excited about raising and educating him (memory and introspection). As a result, it is likelier that John concludes his self-inspection by biasedly judging "I believe that I am healthy" than that John concludes his self-inspection appropriately judging "I have the self-deceived belief that I am healthy". Then, it is likelier that John forms the biased and false$_{rs}$ second-order belief that he believes that he is healthy, failing so to obtain third-person epistemic self-knowledge.

Therefore, because of the expressive properties of self-deceived mental states, it is more challenging to acquire third-person epistemic self-knowledge of a self-deceived mental state than to acquire third-person epistemic self-knowledge of a conscious mental state (e.g., beliefs, desires, intentions, emotions, etc.). However, that doesn't mean that it is not possible to acquire third-person epistemic self-knowledge of self-deceived mental states. In fact, Tom's case (see chapter 1) is an example of *recognized self-deception*, that is, an example of a subject who has third-person self-knowledge of the fact that he is self-deceived. Indeed, Tom has the self-deceived belief that men and women aren't equal (which was formed from deliberation about gender equality), and at the same time, the true$_{rs}$ and warranted second-order belief (i.e., third-person epistemic self-knowledge) that he has the self-deceived belief that men and women aren't equal (which was formed by self-inspecting himself about the attitude that he actually holds about gender equality). Thus, in spite of the difficulties, John could acquire third-person self-knowledge of his self-deceived belief by self-inspecting himself with time and care about the attitude that he really holds about his health condition. In spite of the fact that his self-deceived belief is similar in appearance to the conscious belief that he is healthy and that he is motivationally biased to conclude his self-inspection judging that he is not self-deceived, with time and care John could manage to overcome the difficulties because with time and care John could put his linguistic episodes of expression (e.g., "[I believe that] I am healthy", "I skipped my medical appointment because I was busy", "I enhanced my health insurance because it is important to be cautious in life and I can afford it") into the context of his irrational and conflicting actions (e.g., skipping his medical appointments, enhancing his health insurance, being up for taking a beer with Lydia, driving without fastening his seat belt, borrowing money from Lydia), namely, into the context of the expressive pattern of his self-deceived belief that he is healthy. As a result, John could conclude the process of self-inspection forming the true$_{rs}$

and warranted second-order belief that he has the self-deceived belief that he is healthy, in spite of the difficulties, acquiring so third-person self-knowledge of his self-deceived belief that he is healthy.

However, even if acquiring third-person self-knowledge of the fact that one has a self-deceived mental state might be the first step to overcoming self-deception (for one could try to modify his irrational behaviour when thinking or acting from the first-person perspective), it is not enough to recognize that one is self-deceived about something to overcome self-deception. Due to the motivated and unconscious nature of self-deception, subjects don't usually replace their self-deceived mental states for a conscious mental state (e.g., the self-deceived belief that one is healthy for the conscious belief that one is healthy) straight after discovering that they have a self-deceived mental state. For recognizing or acquiring self-knowledge of the fact that one is self-deceived is not like recognizing or acquiring self-knowledge of the fact that one is mistaken about a non-motivated issue. Thus, after Tom acquires self-knowledge of the fact that he has the self-deceived belief that men and women are not equal, it is expected that Tom still finds himself (by self-inspection) discriminating women somehow every now and then (e.g., by biased judgements, by actions in favour of men, etc.). And after John acquires self-knowledge of the fact that he has the self-deceived belief that he is healthy, it is expected that John still finds himself (by self-inspection) manifesting episodes of his self-deceptive belief that he is healthy every now and then (e.g., avoiding a medical appointment, changing the topic of the conversation, etc.)

In summary, self-deception is explained from the behavioural-expressivist account of Transparency in the following way. When self-deceived subjects answer the question "Do you believe that p?" from the first-person deliberative perspective, they answer with a judgement about whether p that un-self-consciously expresses their self-deceived mental state. As a result, they lack first-person expressive self-knowledge (*knowing how*) both in the sense that they can only express their mental states un-self-consciously and in the sense that it is not possible to have the ability to appropriately express a self-deceived mental state. On the other hand, when self-deceived subjects answer the question "Do you believe that p?" from the third-person perspective, they answer with a self-ascription of attitude that expresses a second-order belief. If this second-order belief is true$_{rs}$ and warranted, they will acquire third-person epistemic self-knowledge of their self-deceived mental states. However, since self-deceived mental states are similar in appearance to conscious mental states (e.g., beliefs) and since they are the result of a bias caused by a motivational state (e.g., the desire to be healthy, anxiety about the possibility of being ill, etc.), it is more difficult to acquire third-person epistemic self-knowledge of the fact that one has a self-deceived mental state than to acquire third-person epistemic self-knowledge of conscious mental states.

Therefore, the behavioural-expressivist account of self-deception understands the cause, the process and the psychological state of self-deception in the following way. Firstly, the *cause* of self-deception is the motivational state (e.g., the desire to be healthy, anxiety about the possibility of being ill, etc.) that causes the bias that affects the collection and assessment of the evidence in the first-person deliberation about whether p. Secondly, the *process* that generates self-deception is the first-person deliberation about whether p. In normal cases, this first-person deliberation is not motivationally biased and it delivers conscious attitudes (e.g., belief, disbelief or suspension). However, in cases of self-deception, this first-person deliberation is motivationally biased in a way that gives rise to the epistemic failure characteristic of self-deception: the subject concludes the deliberation with a judgement about whether p that creates a self-deceived mental state rather than a conscious attitude. Finally, the *psychological state* of self-deception is a *sui generis* and unconscious mental state: a self-deceived mental state.

In what follows, John's example of self-deceived belief is going to be used to describe the three groups of *expressive episodes* that compose the expressive pattern characteristic of self-deceived beliefs:

1) They include a group of expressive episodes that belongs to the expressive pattern of the *self-deceived belief that p* just as much as to the expressive pattern of *anxiety about the possibility of not-p*. For instance, John skips his medical appointments out of anxiety about the possibility of being ill, he got suddenly interested in the afterlife out of anxiety about the possibility of being ill, he enhanced his health insurance more than he can afford out of anxiety about the possibility of being ill, and so on. Thus, these actions are both expressive episodes of John's anxiety about the possibility of being ill and expressive episodes of John's self-deceived belief that he is healthy because they occupy a place in the context of the expressive pattern of these two mental states at once. On the one hand, these actions must be expressive episodes of John's self-deceived belief that he is healthy because without them the irrational and conflicting behaviour (between what the subject says and does) that

is characteristic of self-deception wouldn't take place. On the other hand, these actions are expressive episodes of John's anxiety about the possibility of being ill because John has been anxious about that possibility since the moment that he found the suspicious spot on the back of his shoulder, that is, since before he concluded the first-person deliberation about the character of the spot that gave rise to his self-deceived belief.

2) They include a group of expressive episodes that are *apparently similar* to some of the expressive episodes of the pattern of the *belief that p*. For instance, in John's case of self-deceived belief, there are expressive episodes like saying "[I believe that] I am healthy" or "The spot is just a regular mole. I'm as fit as a fiddle", or like deciding to have a child with the expectation of educating and raising him. These verbal and non-verbal actions are not expressive episodes of John's belief that he is healthy because John doesn't actually believe that he is healthy. Indeed, a subject who firmly *believes* that he is healthy (to the point of deciding to have a child with the expectation of educating and raising him) does not avoid medical appointments or get interested in the afterlife *out of anxiety about the possibility of being ill*. However, those verbal and non-verbal actions must be expressive episodes of John's self-deceived belief that he is healthy because without them the irrational and conflicting behaviour (between what the subject says and does) that is characteristic of self-deception wouldn't take place.

3) They include a group of expressive episodes that are similar in appearance to some of the expressive episodes of the different beliefs that would *rationally explain* the subject's conflicting actions *if she actually believed them*. However, in cases of self-deception, these expressive episodes belong to the pattern of the self-deceived belief that p and not to the pattern of the beliefs that would rationally explain the subject's conflicting actions if she actually believed them.

   For instance, John decides to skip his medical appointments or to enhance his health insurance after deliberating from the first-person perspective about whether he should skip his medical appointment or about whether he should enhance his health insurance, concluding that he shouldn't attend his medical appointment because he is too busy (e.g., "I am skipping my medical appointment today because [I believe that] I have too much work to do to go to the doctor") and that he should enhance his health insurance because being cautious is important in life and he can afford it (e.g., "I am enhancing my health-insurance because [I believe that] being cautious is important in life and [I believe that] I can afford it"). These are expressive episodes of his self-deceived belief that are apparently similar to some expressive episodes of the beliefs that he is too busy to go to the doctor, that it is important in life to be cautious and that he can afford his health insurance; i.e., of the beliefs that would explain John's actions in a rational and justified way if he actually believed them. However, these expressive episodes don't belong to the expressive patterns of those beliefs but only to the expressive pattern of John's self-deceived belief that he is healthy. For John doesn't actually believe that he is too busy to go to the doctor, that it is important in life to be cautious or that he can afford his health insurance. Indeed, that John doesn't believe that he is too busy to go to the doctor in spite of *sincerely* saying "[I believe that] I have too much work to do to go to the doctor" is proved by the fact that he was up for taking a beer with Lydia in the evening; that John doesn't believe that it is important to be cautious in life in spite of *sincerely* saying "[I believe that] it is important to be cautious in life" is proved by the fact that John usually drives without fastening his seat belt; and that John doesn't believe that he can afford the health insurance in spite of *sincerely* saying "[I believe that] I can afford it" is proved by the fact that he recently asked Lydia for money to pay the rent.

As a result, these *beliefs* cannot be the reasons that explain John's actions in a rational and justified way because they cannot be the real reasons that explain why John performed the actions of skipping his medical appointment or enhancing his health insurance (since John doesn't have those beliefs). By contrast, the real reason that explains why John performed these actions is that he has the self-deceived belief that he is healthy (that's why they are irrational and conflicted actions). Indeed, if John hadn't been self-deceived about his health condition, he would have attended his medical appointment (for he wouldn't have considered that he was too busy to go to the doctor) and he wouldn't have enhanced his health insurance more than he can afford (for he wouldn't have considered that being cautious is important in life or that he can afford it). As a result, expressive episodes like saying "[I believe that] I have too much work to do" or "[I believe that] being cautious is important in life and I can afford it" are episodes of *reason-confabulation* that express John's self-deceived belief that he is healthy. More precisely, they are expressive episodes of John's self-deceived belief that he is healthy that consist in falsely confabulating the beliefs that would have explained John's actions in a *rational way* (if John had believed them, of course).

Therefore, the expressive pattern of self-deceived beliefs can be described by saying that it is composed of those three groups of expressive episodes. In the next section, it is going to be argued that, unlike the accounts of self-deception currently available, the behavioural-expressivist account of self-deception proposed here meets all the desiderata of a good account of self-deception.

3.6 Accounting for the desiderata of self-deception

In this section, it is going to be argued that the behavioural-expressivist account of self-deception proposed here satisfies all the desiderata of a good account of self-deception. Since the accounts of self-deception currently available in the literature are not able to explain all the desiderata of the phenomenon, it will be concluded that the behavioural-expressivist account is the best account of self-deception currently available. Let's see how the behavioural-expressivist account of self-deception explains all the desiderata in order:

1) It explains the *irrational conflict* between what the subject sincerely says and how she acts characteristic of self-deception because it claims that self-deceived mental states are unconscious mental states. So when self-deceivers linguistically express their self-deceived mental states, their expressive episodes are similar in appearance to the expressive episodes characteristic of some conscious mental states (e.g., "[I believe that] I am healthy", "[I believe that] I don't have time to go to the doctor today", "[I believe that] I can afford my health insurance"), but they act in an irrationally conflicted way with what they say (e.g., skipping medical appointments or avoiding talks about medical issues, being up for having a beer the same day as

the medical appointment was scheduled or borrowing some money from Lydia to pay the rent). That this conflict between what John says and does is irrational is proved by the fact that self-deceivers cannot express their self-deceived mental states with their true$_{\text{nrs}}$ first-person avowals (e.g., “I am self-deceived about the fact that I am healthy”).

2) It explains why self-deceived subjects are considered to suffer from a lack of self-knowledge and to have some kind of *unconscious* mental state. On the one hand, the intuition that self-deceivers have some kind of unconscious mental state is explained because being self-deceived consists in having a certain kind of unconscious mental state (i.e., a self-deceived mental state). On the other hand, the intuition that self-deceivers suffer from a lack of self-knowledge is explained in the following way. When John answers Lydia’s question “Do you believe that you are healthy?” from the first-person deliberative perspective, he self-unconsciously expresses his self-deceived belief that he is healthy with the judgement that he is healthy because he expresses his self-deceived belief as if he were expressing a conscious belief (i.e., without knowing what he is up to). Indeed, there isn’t any sense in which John could have first-person expressive self-knowledge when he expresses his self-deceived belief because self-deceived mental states can only be expressed un-self-consciously (i.e., without knowing what one is up to), and so, it is not possible to have the ability to appropriately express an un-self-conscious mental state. By contrast, when John answers Lydia’s question from the third-person self-inspective perspective, he answers with a self-ascription of attitude that expresses a second-order belief on the basis of evidence about his mental states. If the second-order belief is true$_{\text{rs}}$ and warranted, he will acquire third-person epistemic self-knowledge. However, in cases of self-deception, it is difficult to acquire third-person epistemic self-knowledge of one’s own self-deceived mental state because the expressive episodes of self-deception are similar in appearance to the expressive episodes of other conscious mental states and because one is motivationally biased to conclude that one is not self-deceived.

From the third-person self-inspective perspective, it is explained as well why people say things like “I was fooling myself” or “Deep down, I knew the truth all along” after overcoming self-deception. On the one hand, in regard to “I was fooling myself", people often say so when they overcome self-deception because self-deceived mental states always have expressive episodes that are false$_{nrs}$ in their patterns of expression (e.g., the false$_{nrs}$ avowal "I believe that p"). Since self-deceived mental states are unconscious mental states, false$_{nrs}$ avowals are the only explicit linguistic expressive episodes that belong to their expressive pattern. On the other hand, in regard to "Deep down, I knew the truth all along", people often say so when they overcome self-deception because, when self-inspecting themselves from the third-person self-inspective perspective, they realize that they had the evidence to make a non-motivated judgement from the first-person perspective, and so, that they should have formed a conscious attitude about p rather than a self-deceived mental state about p.

3) It explains why self-deception has a high degree of *persistence* against criticism based on evidence insofar as it explains the subject's *rationalizations* of the evidence behind the phenomenon of persistence. These rationalizations of the evidence occur when self-deceivers *deliberate* from the first-person perspective about whether p (e.g., about whether one is healthy) or about whether they should perform certain actions (e.g., about whether one should attend his medical appointment or about whether one should enhance his health insurance).

   Firstly, self-deceivers rationalize the evidence about p because *some* of the expressive episodes of self-deceived mental states are identical to the biasing mechanisms described by Mele (2001, pp. 25-27), and so, *some* of the expressive episodes of self-deceived mental states implicitly bias the collection and assessment of the evidence about p that subjects might use when deliberating from the first-person perspective to answer the question "Do you believe that p?". For instance, John skips his medical appointments out of anxiety about the possibility of being ill, and in doing so, he doesn't collect evidence where he could find that he is ill against his own epistemic norms (i.e., *selective evidence-gathering*). When Lydia points out to John that the spot deserves to be checked by a professional, John disqualifies Lydia's skill to tell when a spot deserves to be checked out of anxiety about the possibility of being ill, and in doing so, he disqualifies evidence (i.e., Lydia's opinion) that might point to the fact that he is ill against his own epistemic norms (i.e., *negative misinterpretation*). Or when John deliberates about the character of the spot, he focus more on the colour than on the shape out of anxiety about the possibility of being ill because the shape look more suspicious to him, and in doing so, he avoids focusing on evidence that might point to the fact that he is ill against his own epistemic norms (i.e., s*elective focusing or attending*). As a result of these rationalizations, John's self-deceived belief that he is healthy tends to be persistent against criticism based on evidence, at least more persistent than regular non-motivated mistakes, when he deliberates from the first-person perspective to answer the question "Do you believe that you are healthy?".

Secondly, when self-deceivers answer the question "Should you do such and such?" (e.g., "Should you enhance your health insurance?" or "Should you skip your medical appointment?") deliberating from the first-person perspective about whether they should perform certain actions, they rationalize the evidence about what they should do because they are motivationally biased to conclude the deliberation with an episode of belief-confabulation that expresses their self-deceived mental state (e.g., "[I believe that] I have too much work to do to go to the doctor today" or "[I believe that] I am going to enhance my health-insurance because being cautious is important in life and I can afford it"). Particularly, self-deceivers are motivationally biased to conclude their deliberation about whether they should perform a particular action with an episode of belief-confabulation consisting in the same judgement that would have given rise to the belief that would have justified their action in a rational way (if they had actually believed it). For instance, when deliberating about whether he should attend his medical appointment today, John is motivationally biased to conclude that he shouldn't because he finds distress in thinking about the importance of attending his medical appointment and relief in thinking about staying at home finishing some work. As a result, John overestimates the time that he is going to need in order to finish the work and he hastily and biasedly concludes "[I believe that] I am too busy to go to the doctor today". This judgement is an expressive episode of his self-deceived belief that he is healthy that consists in confabulating the belief that would have explained his action of skipping the medical appointment in a rational way if he had believed it (for, remember, John didn't actually believe that he was busy, as it is proved by the fact that he was up for taking a beer with Lydia that day).

4) It accounts for why subjects are *epistemically responsible* for being self-deceived because it claims that self-deceived mental states are reason-sensitive[48] mental states. Self-deceived mental states are formed by a motivationally biased first-person deliberation about whether p (which would answer the transparent question "Do you believe that p?") and they can be overcome and replaced by a conscious mental state also by first-person deliberation about whether p (which would answer the transparent question "Do you believe that p?").

Indeed, self-deception can be overcome by first-person deliberation about whether p because there is a threshold of evidence about whether p that a self-deceived subject can acquire to replace her self-deceptive mental state by a conscious attitude (e.g., belief, desire, intention…). For instance, if John ended up attending a medical appointment and if the doctor managed to make him undertake thorough medical tests, John would be able to overcome self-deception by finding out the results of his medical tests (assuming that they are conclusive). For, on the basis of such overwhelming evidence, he would end up replacing his *self-deceived belief* that he is healthy for the *belief* that he is healthy (if the results are negative) or for the *belief* that he is ill (if the results are positive). If no counterfactual situation like that could occur in John's case, John wouldn't be self-deceived about the fact that he is healthy because there wouldn't be any threshold of evidence that John could reach to overcome self-deception, and so, he wouldn't be epistemically responsible for having his irrational mental state (i.e., he would be in the same situation as those schizophrenic patients who say they believe that their caregivers are trying to kill them without refusing the food that their caregivers provide them with). Therefore, self-deceived subjects are epistemically responsible for being self-deceived because there is a threshold of evidence that they could reach to overcome self-deception, and so, self-deceived mental states are reason-responsive or reason-sensitive.

[48] Again, see McHugh (2013) for an account of epistemic responsibility based on the conditions of 1) doxastic agency and 2) reason-responsiveness.

Thus, the behavioural-expressivist account is the only account of self-deception, among the accounts of self-deception currently available in the literature, that explains all the desiderata that a good account of self-deception should be able to explain.

3.7 Remapping the terrain: wishful thinking, self-deception and delusion

In this section, it is going to be argued that the behavioural-expressivist account of self-deception has the additional advantage of being able to appropriately distinguish and classify some phenomena which are considered to be different in the literature (either in kind or in degree) even if they are closely related to each other. The phenomena of wishful thinking, self-deception and delusion. The reason why the behavioural-expressivist account of self-deception is able to appropriately distinguish between these phenomena is that it claims that self-deception is an unconscious mental state, so that it is different from the conscious mental states involved in wishful thinking (e.g., belief), and that it claims that self-deception is a reason-responsive mental state, so that it is different from delusions because these are not-reason-responsive. Since wishful thinking and self-deception are the result (i.e., psychological state) of a motivated bias (i.e., a process), let’s differentiate between motivated and non-motivated biases first.

Firstly, an *implicit bias* occurs when there is something (e.g., the way in which our cognitive system is made or a mental state) that, unbeknown to the subject, affects her collection and assessment of the evidence about p in a way that goes against her own epistemic norms. The cause of an implicit bias can be a motivational state or not. On the one hand, an example of non-motivated implicit bias is the *confirmation bias* (Beattie & Baron, 1988; Mele, 2001): subjects tend to test hypothesis or beliefs in a way that favours the confirmation of the hypothesis or belief (against their epistemic norms) because they unknowingly tend to take into account evidence confirming the hypothesis or belief rather than evidence falsifying it. For instance, when a group of people are set to test the hypothesis that someone’s facial expression is an expression of anger and another group of people are set to test the hypothesis that someone’s facial expression is an expression of happiness, the former group tend to conclude that it is a facial expression of anger and the latter that it is a facial expression of happiness, in spite of the fact that the shown facial expression was the same in both cases (Mele, 2001; Trope, Gervey & Liberman, 1997). By contrast, an example of motivated implicit bias is John's deliberation about whether the spot on the back of his shoulder is malignant or not. Since John wants to be healthy (i.e., motivational state), he finds distress in focusing on the shape of the spot (which he finds suspicious) and relief in focusing on the colour of the spot (which he finds similar to the colour of a regular mole). As a result, he ends up judging that the spot is just a regular mole because, against his own epistemic norms and unknowingly to him, he overweighs the evidence that has to do with the colour and underweights the evidence that has to do with the shape.

Secondly, it has been claimed (e.g., Mele, 2001; Szabados, 1973) that the difference between self-deception and wishful thinking is that in cases of self-deception the subject falsely judges that p when she has evidence *against* p, while in cases of wishful thinking the subject falsely judges p without having enough evidence *either for p or against p*. Thus, the difference in regard to the evidence available to the subject is supposed to explain why intuitively self-deception and wishful thinking seem to be different phenomena (either in kind or in degree) insofar as self-deception seems more irrational and conflicted than wishful thinking. However, in the behavioural-expressivist account of self-deception defended here, the phenomena of motivated bias, wishful thinking and self-deception are considered different kinds of phenomena and they are distinguished in the following way. Both *wishful thinking* and *self-deception* are considered to be the result (i.e., psychological state) of a *motivated implicit bias* in the process of first-person deliberation about whether p (for both wishful thinking and self-deception are the result of the subject unknowingly collecting and assessing the evidence against her own epistemic norms because of a motivational state). However, wishful thinking occurs when the output of the motivationally biased deliberation is a *conscious attitude* (i.e., belief, desire or intention), whereas that self-deception occurs when the output of the motivationally biased deliberation is an (unconscious) *self-deceived mental state*.

On the one hand, in cases of wishful thinking, the subject is motivationally biased when he deliberates about whether p, but when he concludes the deliberation with a biased judgement about whether p, no expressive failure occurs so that she forms the conscious attitude that corresponds to the judgement. As a result, the subject has first-person expressive self-knowledge of her conscious attitude, even if it is the result of wishful thinking, and she won't act in the irrational conflicted way characteristic of self-deception. For instance, imagine a subject who is motivationally biased when she deliberates about whether the spot on the back of her shoulder is malignant, but she concludes the deliberation biasedly judging that it is a regular mole and forming the (conscious) belief that it is a regular mole (so that no expressive failure occurs). This is a case of wishful thinking, and so, the subject won't manifest the irrational conflicting behaviour that is characteristic of self-deception because the subject actually believes that p (e.g., she won't skip her medical appointments out of anxiety about the possibility of being ill, she won't enhance her health insurance more than she should out of anxiety about the possibility of being ill… and she won't try to rationalize those conflicting actions with episodes of belief-confabulation). As a result, the subject has the biased (conscious) belief that she is healthy. On the other hand, in cases of self-deception, the subject is motivationally biased when she deliberates about whether p in a way that an expressive failure occurs when she concludes the deliberation. When she biasedly judges that p, she forms the (unconscious) self-deceived mental state that p rather than a conscious attitude about p (i.e., belief, desire or intention). John's case is an example of a subject who biasedly deliberates about whether the spot on the back of his shoulder is malignant and concludes the deliberation with a biased judgement that gives rise to the (unconscious) self-deceived belief that he is healthy rather than to the (conscious) belief that he is healthy. As a result, John will manifest the irrational and conflicting behaviour that is characteristic of having a self-deceived belief and he will suffer from a lack of first-person self-knowledge about that mental state because it cannot be self-consciously expressed.

Thirdly, it has been discussed whether motivational biases are involved in cases of *delusions* or not (e.g., Bayne & Fernández, 2008). However, according to the behavioural-expressivist account defended here, regardless of whether there is a motivated bias involved in some cases of delusion or not, self-delusions are a *sui generis* kind (Egan, 2008; Tumulty, 2012) of unconscious mental states (i.e., the true$_{\text{nrs}}$ avowal "I am deluded about p" is not an expressive episode of their expressive pattern) which differ from self-deceived mental states because subjects are not epistemically responsible for having them. As a result, deluded subjects also suffer from lack of first-person self-knowledge about their self-deluded mental state (i.e., they cannot be self-consciously expressed), and so, they also manifest an irrational conflict between what they sincerely say and how they act. This chapter has already mentioned the case of a delusional schizophrenic patient who sincerely claim that her caretakers were trying to kill her but who willingly ate the food that they provided her with. Let's see the example of *Capgras delusion* now, which appears in some subjects who have suffered brain damage. Subjects with Capgras delusion (Stone & Young, 1997) sincerely say to believe that one or more of their close relatives have been replaced by an identical counterfeit (e.g., a clone, a robot, a Martian…). However, they manifest the same kind of conflicting and irrational behaviour that it is characteristic of self-deception: even if some of their actions are coherent with what they claim to believe (e.g., they might refuse to sleep with their wife or they might order the relative to leave the house), other actions are irrationally conflicted with what they say to believe (e.g., they don't show much interest in what happened to the real relative and they don't search for them).

The fact that delusional mental states and self-deceived mental states share the same kind of irrational conflict between what the subject sincerely says and how she acts is explained because both delusional mental states and self-deceived mental states are different kinds of unconscious mental states (i.e., neither the true$_{nrs}$ avowal "I am self-deceived about the fact that I am healthy" nor the true$_{nrs}$ avowal "I am deluded about the fact that my wife is a clone" are expressive episodes of their respective expressive patterns), and so, they cannot be self-consciously expressed by the subject. Delusional mental states and self-deceived mental states, though, are *different* kinds of unconscious mental states because of the following fact: while self-deceived mental states are reason-sensitive because there is a threshold of evidence that can be acquired by the subject to overcome self-deception, mental states of delusion are not reason-sensitive (Szabados, 1985, p. 160; DSM-V, 2000, p. 819) because, insofar as delusions are caused (Silva et al., 1998; Soyka, 1995; Soyka & Schmidt, 2011) by a psychiatric disorder (e.g., schizophrenia) or physiological damage (e.g., brain damage), there is no threshold of evidence that deluded subjects can reach to overcome their delusions. Indeed, no matter how much evidence is given to subjects who suffer from Capgras delusion about the fact that their relatives haven't been replaced by a counterfeit, if they ever overcome their delusional state, it won't be because of that evidence. As a result, subjects are epistemically responsible for having a self-deceived mental state but not for having a delusional mental state.

In summary, in this chapter has it been argued that from the behavioural-expressivist account of Transparency follows the best account currently available of self-deception. In order to do that, the relational and the non-relational senses of truth have been explicated and the concept of unconscious mental state has been characterized on that basis. Then, it has been argued that self-deception is an unconscious mental state that is formed by a motivated first-person deliberation and that involves both a lack of first-person expressive self-knowledge (*knowing how*) and difficulties to acquire third-person epistemic self-knowledge (*knowing that*). Self-deceived subjects lack first-person expressive self-knowledge of their self-deceived mental states because they cannot be self-consciously expressed (e.g., when answering the transparent question “Do you believe that p?” from the first-person deliberative perspective) insofar as they can only be expressed with false$_{nrs}$ avowals. On the other hand, self-deceived subjects have difficulties to acquire third-person epistemic self-knowledge (i.e., difficulties to answer the question “Do you believe that p? from the third-person self-inspective perspective with a true self-ascription of attitude) because the expressive episodes of self-deceived mental states are similar in appearance to expressive episodes of conscious mental states and because self-deceivers are motivationally biased to conclude that they are not self-deceived.

In the next chapter, it is going to be argued that from the behavioural-expressivist account of Transparency follows the best explanation of Moore’s paradox as well.

4. Moore’s Paradox

Usually, there is a certain kind of irrationality in asserting first-person present sentences with the logical form “p, but I don’t believe that p” (e.g., “It is raining, but I don’t believe so”) and “p, but I believe that not-p” (e.g., “It is raining, but I believe that it isn’t”). The irrationality of asserting sentences with such a logical form arises only when they are asserted in the *first-person present tense*. For, on the one hand, there isn’t any kind of irrationality in asserting in the past tense “p, but I didn’t believe that p” (what can be said to attribute *lack of belief* about p to oneself in the past) and “p, but I believed that not-p” (what can be said to attribute a *mistaken belief* to oneself in the past). And, on the other hand, there isn’t any kind of irrationality in asserting the second-person sentences or the third-person sentences “p, but you/she don’t/doesn’t believe that p” (what can be said to attribute *lack of belief* about whether p to others) and “p, but you/she believe/s that not-p” (what can be said to attribute a *mistaken belief* to others). Hereafter, first-person present sentences with the logical form “p, but I don’t believe that p” or “p, but I believe that not-p” will be called “Moore’s sentences” for short.

It is considered perplexing that Moore’s sentences are irrational to assert because they are supposed to describe very possible situations, and so, they are not considered to be *self-contradictory*. Indeed, we use sentences like “p, but I didn’t believe that p” or “p, but I believed that not-p” to describe cases in which one didn’t have any belief about p or had a mistaken belief about p, and we use sentences like “p, but you/she don’t/doesn’t believe that p” and “p, but you/she believe/s that not-p” to describe cases in which others don’t have any belief about p or have a mistaken belief about p. Hence, it seems that Moore’s sentences have possible truth-conditions and that those truth-conditions have to do with whether I lack a belief about something or whether I have a mistaken belief about something (which is a possible —and indeed very likely— psychological fact about myself). Then, if Moore's sentences are not self-contradictory, how is it that we consider it irrational to use a Moore's sentence to attribute lack of belief or error to oneself in the present tense? This paradox was first pointed out by Moore (1993) and it was labelled as "Moore's paradox" later on by Wittgenstein (1953).

Depending on the kind of irrationality that is thought to be involved in asserting a Moore's sentence, four kinds of accounts of Moore's paradox can be distinguished. *Pragmatic accounts* (Baldwin, 1990; Moore, 1993; Rosenthal, 2005; Hamilton, 2014) claim that Moore's paradox arises because Moore's sentences don't have appropriate conditions of assertion. *Psychological accounts* (Baldwin, 2007, 1990; Coliva, 2016; Shoemaker, 1996; Williams, 2006) claim that Moore's paradox arises because to assert a Moore's sentence involves an inconsistency among the subject's mental states. *Epistemic accounts* (Fernández, 2005, 2013; Moran, 1997, 2001) claim that Moore's paradox arises because to assert a Moore's sentence involves an epistemic failure that causes the subject's lack of first-person epistemic self-knowledge. And *semantic accounts* (Heal, 1994; Linville & Ring, 1991) claim that Moore's paradox arises because to assert a Moore's sentence, in spite of appearances, involves some kind of contradiction or contradiction-like.

The aim of this chapter is to argue that from the behavioural-expressivist account of Transparency follows a *semantic account* of Moore's paradox and that the kind of semantic account of Moore's paradox that follows from the behavioural-expressivist account of Transparency is able to offer the best explanation of the phenomenon of Moore's paradox among the accounts currently available in the literature. In order to do that, the argument will go as follows. Firstly, the desiderata that every good account of Moore's paradox should explain will be spelled out. Secondly, the different types of accounts of Moore's paradox will be explicated and it will be argued that all of them fail to explain at least one of the desiderata. And, finally, it will be argued that the semantic account that follows from the behavioural-expressivist account of Transparency is the best explanation of Moore's paradox because it is the only account that manages to explain all the desiderata.

4.1 The desiderata of an account of Moore's paradox

Four intuitions regarding Moore's paradox usually appear in the discussion of the phenomenon. Each one of these intuitions can be considered a *desideratum* that every good account of Moore's paradox should be able to explain. The desiderata of Moore's paradox that have been pointed out in the literature are the following four:

1) *Moore's paradox arises when the content of Moore's sentences is irrational to assent*. One qualification is needed to clarify this desideratum: that Moore's paradox arises only when the two parts of the Moore's sentence (i.e., both "p" and "I don't believe that p" or "I believe that not-p") are assented from the *deliberative first-person perspective* (regardless of whether the deliberative first-person perspective is understood in an epistemic or in a behavioural-expressivist way) because only under that condition the content of a Moore sentence is irrational to assent.

   Indeed, Moore's paradox arises only when a Moore's sentence is assented from the deliberative first-person perspective because, as soon as the third-person process of self-inspection is involved in the assent of any of the parts of a Moore's sentence (i.e., either "p" or "I don't believe that p"/"I believe that not-p), the irrationality of assenting to that Moore's sentence disappears (and so, Moore's paradox doesn't arise). Thus, Moore's paradox is an *only* first-person phenomenon, just like Transparency. For instance, Tom's assent to "Men and women are equal, but I don't believe so" is not an instance of Moore's paradox because "I don't believe so" is the result of a third-person process of self-inspection, and hence, it is not irrational for Tom to assent so. Similarly, assenting to "He is dead. I can't believe it"[49] is not an instance of Moore's paradox when the subject assents to "I can't believe it" (e.g., in the edge of the sudden death of a close friend) as a result of a third-person process of self-inspection because it is not irrational for him to assent so under that

[49] Baldwin (1990) understands that the assertion "He is dead. I can't believe it" is an instance of Moore's paradox. In order to understand the sentence as an instance of Moore's paradox, it is necessary to understand that "I can't believe it" is not the result of third-person self-inspection but of first-person deliberation (just like "He is dead").

condition. To see the difference with actual cases of Moore's paradox, compare these examples with a case in which the assent to "It is raining, but I don't believe so" is made from a full first-person deliberative perspective. In this case, it is clear that the assent is irrational *to make*, and so, that the assent is an instance of Moore's paradox.

The intuition that Moore's paradox is an only first-person phenomenon is usually pointed out in the literature in the following way. It is claimed that the irrationality of Moore's paradox doesn't arise merely because of the fact that the subject is in an irrational cognitive state that she expresses or describes (or both) by assenting to a Moore's sentence; rather, the irrationality of Moore's paradox is supposed to arise because Moore's sentences are irrational *to assent* from the first-person deliberative perspective; i.e., because *the act* of assenting to a Moore's sentence from the first-person deliberative perspective is irrational in itself. (Coliva, 2016, p. 254; Fernández, 2005, p. 541; Heal, 1994, p.11).

2) *Moore's paradox should be explained both in the case of "p, but I don't believe that p" and in the case of "p, but I believe that not-p".* For there is a conceptual difference between the Moore's sentence "p, but I don't believe that p" and the Moore's sentence "p, but I believe that not-p" (Coliva, 2016; Fernández, 2005, 2013; Heal, 1994; Moran, 1997, 2001; Williams, 2006). This conceptual difference is due to the fact that "I don't believe that p" can be used to make a self-ascription of *lack of belief* about whether p, while "I believe that not-p" cannot be used to make a self-ascription of lack of belief about whether p.

The way in which this conceptual difference is usually fleshed out in the literature is similar to the way in which epistemic accounts of Transparency understand first-person avowals. So, "I believe that not-p" is considered to be a self-ascription of the belief that not-p both when it is an answer to the deliberative question "Do you believe that p?" (first-person avowal) and when it is an answer to the self-ascriptive question "Do you believe that p?" (third-person avowal); and "I don't believe that p" is considered to be (at least sometimes[50]) a self-ascription of lack of belief about p both when it is an answer to the deliberative question "Do you believe that p?" (first-person avowal) and when it is an answer to the self-ascriptive question "Do you believe that p?" (third-person avowal). However, this interpretation of the conceptual difference between "I believe that not-p" and "I don't believe that p" is not mandatory. Instead, in line with the behavioural-expressivist account of Transparency, the following understanding of the conceptual difference between the two Moore's sentences can be offered. On the one hand, "I believe that not-p" can be used both to express the belief that not-p by judging that not-p (when it is a first-person avowal issued as an answer to the deliberative question "Do you believe that p?") and to make a self-ascription of the belief that not-p (when it is a third-person avowal issued as an answer to the self-inspective question "Do you believe that p?"). On the other hand, "I don't believe that p" can be used both to express the subject's lack of belief about whether p without judging either p or not-p (when it is a first-person avowal of suspension of judgement issued as an answer to the deliberative question "Do you believe that p?") and to make a self-ascription of lack of belief about whether p (when it is a third-person avowal issued as an answer to the self-inspective question "Do you believe that p?").

3) *Moore's sentences can be true$_{rs}$ because they can describe possible situations*. Particularly, it has been argued (Fernández, 2005, 2013; Moran, 1997, 2001; Shoemaker, 1996) that Moore's sentences can describe cases in which p is the case but I don't believe that p (i.e., "p, but I don't believe that p") or cases in which p is the case but I believe that not-p (i.e., "p, but I believe that not-p"). Tom's assertion "Men and women are equal, but I don't believe so" is an example of a Moore's sentence used to describe a case in which men and women are equal but the speaking subject doesn't believe so because he acts in a way that it is incompatible with believing that men and women are equal.

4) *Moore's paradox should be explained both in cases in which a Moore's sentence is linguistically asserted and in cases in which a Moore's sentence is silently judged*

[50] Notice that the sentence "I don't believe that p" can sometimes be used in the same sense as "I believe that not-p". However, "I don't believe that p" has its own differentiated use on other occasions.

> *in thought*. Indeed, even if Moore's paradox is usually formulated by attending to the irrationality that arises when *asserting* a Moore's sentence, it is usually considered that a good account of Moore's paradox should explain as well why it is also irrational to *silently think* a Moore's sentence (Baldwin, 2007; Fernández, 2005, 2013; Moran, 1997, 2001; Shoemaker, 1996; Heal, 1994; Coliva, 2016; Williams, 2006), that is, why it is also irrational to make a judgement with the content of a Moore's sentence in thought and without pronouncing anything aloud.

Therefore, a good account of Moore's paradox should be able to explain why (1) sometimes it is irrational (4) to judge in thought or to assert aloud (2) the sentences "p, but I don't believe that p" and "p, but I believe that not-p" in spite of the fact that (3) such sentences can have possible truth$_{\text{rs}}$-conditions. The main accounts of Moore's paradox available in the literature are going to be explicated in the following sections, and it will be concluded that they all fail to account for the phenomenon of Moore's paradox because they all fail to explain appropriately at least one of the desiderata of a good account of Moore's paradox.

4.2 Pragmatic accounts of Moore's paradox

Pragmatic accounts (Baldwin, 1990; Moore, 1993; Rosenthal, 2005; Hamilton, 2014) claim that Moore's paradox arises because Moore's sentences don't have appropriate *assertion-conditions* in those conversational contexts in which they are irrational to assert (i.e., in those conversational contexts in which Moore's sentences are instances of Moore's paradox)[51], even if they always have appropriate *truth-conditions* (that's why they are not self-contradictions).

[51] Strictly speaking, pragmatic accounts don't explicitly distinguish between conversational context in which the assertion of a Moore's sentence is an instance of Moore's paradox because it is irrational to assert (i.e., when the utterance is made as a result of a first-person deliberation about whether p) and conversational contexts in which the assertion of a Moore's sentence is not an instance of Moore's paradox because it is not irrational to assert (i.e., when the utterance is made as a result of a third-person process of self-inspection). However, since pragmatic accounts are perfectly compatible with this true distinction between conversational contexts, I prefer to understand them in a charitable way and suppose that they assume this distinction even if they don't explicitly formulate it. By doing so, I want to avoid criticizing pragmatic accounts on the basis of the superficial reason that they don't distinguish between the two contexts (for pragmatic accounts are perfectly compatible with distinguishing between them).

Then, Moore's sentences are irrational to assert when they don't have appropriate assertion-conditions even if their truth-conditions can be satisfied.

Indeed, the sentence "p, but I don't believe that p" is considered to be true about me whenever p is the case but I don't have any belief about p, and the sentence "p, but I believe that not-p" is considered to be true about me whenever p is the case but I have a mistaken belief about p. Since it is obvious that I might not have a belief about something that is the case or that I might have a mistaken belief about something that is the case (in fact, it is extremely likely that that is so), the sentences "p, but I don't believe that p" and "p, but I believe that not-p" can be true about me right now (in fact, it is extremely likely that they are true about me right now). So, if such sentences are irrational to assert in first-person deliberative contexts, the irrationality can only arise because of a problem in their assertion-conditions and not in their truth-conditions.

Moore (1993) himself offered an account of Moore's paradox in terms of inappropriate assertion-conditions. According to Moore, to assert "p, but I don't believe that p" and "p, but I believe that not-p" is irrational because to assert "p" *implies* both that one believes that p and that one doesn't believe that not-p. The nature of this implication is inductive. For it is based on the fact that we learn from experience that people don't usually lie when they assert something about the world. Thus, on the one hand, to assert "p, but I don't believe that p" is irrational because the subject asserts that *she doesn't believe that p*, and at the same time, she *implies* that *she believes that p* by asserting "p". On the other hand, to assert "p, but I believe that not-p" is irrational because the subject asserts that *she believes that not-p*, and at the same time, she implies that *she doesn't believe that not-p* by asserting "p". Since these implications don't conflict neither with "I believed that p" (in the past) nor with "She believes that p" (in third-person), sentences like "p, but I didn't believe that p" or "p, but she doesn't believe that p" are not irrational to assert. Therefore, the irrationality of asserting a Moore's sentence has nothing to do with their truth-conditions (for they can be true even when they are irrational to assert) but with the fact that they don't have appropriate conditions of assertion insofar as asserting "p" implies both that the subject believes that p and that the subject doesn't believe that not-p.

However, Moore's account of the paradox suffers from a problem that is independent of its capacity to explain the desiderata of a good account of Moore's paradox. The problem is the following: Moore's sentences are irrational to assert even when the person who utters them is a well-known pathological liar so that no implication from asserting "p" to believing that p and to disbelieving that not-p should take place. For, remember, that implication is considered to be based on experience and induction, and nobody would make that induction in the case of a subject who is well-known to be a pathological liar (Baldwin, 1990).

Baldwin (1990), by contrast, offers a pragmatic account of Moore's paradox based on the Gricean notion of meaning. Baldwin considers, just like Moore, that to assert a Moore's sentence is irrational because to assert "p" implies something about the subject's beliefs. However, he thinks that this implication arises, not because of an inductive inference from the fact that people don't usually lie, but because of the fact that the intention of informing one's audience through the recognition that one has the intention of informing one's audience is constitutive of the speech act of assertion. Indeed, according to Grice (1957), it is constitutive of the speech act of asserting "p" that the speaker intends that the audience believes p on the basis of the recognition that *that* is precisely the speaker's intention. In order to explain Moore's paradox, Baldwin adds the claim that the speaker's intention that the audience believes that p on the basis of the recognition that *that* is precisely the speaker's intention includes the intention to be taken as *believing* what one asserts. For nobody would form the belief that p on the basis of the recognition that the speaker intends one to believe that p if she didn't take the speaker as already believing that p.

Thus, Moore's paradox is explained as follows. On the one hand, it is irrational to assert "p, but I don't believe that p" because the assertion of "p" involves the speaker's intention to be taken by her audience as already believing that p, which contradicts what is asserted in the second-half of the Moore's sentence: "I don't believe that p". On the other hand, it is irrational to assert "p, but I believe that not-p" because the assertion of "p" involves the speaker's intention to be taken by her audience as already believing that p, belief that it is inconsistent with the belief that the speaker asserts to have in the second half of the Moore's sentence: "I believe that not-p". Therefore, the apparent intention of the speaker when asserting either "p, but I don't believe that p" or "p, but I believe that not-p" seems to involve the need for her audience to attribute to her either contradictory attitudes (i.e., that she believes that p and that she doesn't believe that p) or inconsistent beliefs (that she believes that p and that she believes that not-p). By contrast, no apparent intention of the speaker seems to involve the need for her audience to attribute to her inconsistent or contradictory attitudes when she asserts either "p, but I didn't believe that p" or "p, but she doesn't believe that p".

Moreover, since the speaker's intention is constitutive of the meaning of the speech act of assertion, Baldwin's account is supposed to appropriately predict that Moore's sentences are irrational to assert even when the speaker is a well-known liar. For, even when the audience recognizes that the speaker is lying, the speech act requires, in order to be understood, that the speaker intends to be recognized by her audience as having the intention to truly inform them of something that she believes.

However, regardless of whether Baldwin's account explains all the desiderata of Moore's paradox or not, it suffers from an independent problem that was pointed out later by Baldwin himself and that led him to abandon his pragmatic account definitively (Baldwin, 2007). The problem is that there are situations in which Moore's paradox arises and the speaker seems to assert "p" without the intention for her audience to form the belief that p on the basis of the assertion. For instance, imagine that someone is taking an oral exam and she says "p, but I don't believe that p". It is obvious that the examiners already know the right answers to the questions and that the intention of the examinee is not to make the examiners believe the answers (for they already believe the right answers) but to make the examiners believe that she knows the answers. However, in spite of that, the examinee's assertion "p, but I don't believe that p" is irrational to make, and so, it gives rise to an instance of Moore's paradox also in that conversational context.

Furthermore, regardless of the theoretical problems that internally arise in Moore's and Baldwin's accounts, both accounts are unable to explain one *desideratum* of Moore's paradox: 4) why Moore's paradox arises, not only when asserting a Moore's sentence, but also when judging the content of a Moore's sentence silently in thought. Indeed, when one silently judges in thought "p, but I don't believe that p" or "p, but I believe that not-p", one doesn't assert anything, but the irrationality characteristic of Moore's paradox is as present as when one asserts them aloud. How could Moore and Baldwin explain this aspect of the paradox? It seems that they can't. For Moore and Baldwin claim that Moore's paradox arises because of an implication (either based on an inductive inference or on the speaker's intentions) made on the basis of the subject's *assertion*, and when a Moore's sentence is judged silently in though, there is no assertion to begin with and no implication from it can be triggered.

Rosenthal offers a different pragmatic account of Moore's paradox that avoids the internal problems of Moore's and Baldwin's accounts, while promising to account for the fact that Moore's paradox arises, not only when asserting a Moore's sentence, but also when silently judging its content. Rosenthal argues that Moore's paradox arises because the assertions "p" and "I believe that p" have similar *assertion-conditions* (at least in conversational contexts in which Moore's paradox arises) in spite of the fact that they have different *truth-conditions*. In order to justify this claim, Rosenthal distinguishes between *expressing* and *reporting* a mental state. To express a mental state and to report a mental state are two different ways of conveying the information than one has a mental state. On the one hand, a subject expresses a mental state by a speech act when i) the speech act has the same content as the mental state and ii) the speech act has the illocutionary force that corresponds in speech to that particular kind of mental state. For instance, the utterance "It is raining" expresses the belief that it is raining because it has the content "It is raining" and the illocutionary force of an assertion; or the utterances "Chocolate!" and "To have some chocolate would be nice" express the desire to eat chocolate because they have the content "Eating chocolate" and the illocutionary force of wanting or desiring something. On the other hand, a subject reports a mental state by a speech act when i) the speech act has a different content than the mental state and ii) the illocutionary force of the speech act is an assertion (reports are always assertions). For instance, the assertions "I believe that it is raining" and "I want chocolate" report the speaker's belief that it is raining and the speaker's desire to eat chocolate because their contents are the *belief* that it is raining (i.e., "I believe that it is raining") and the *desire* to eat chocolate (i.e., "I want to eat chocolate") and because they have the illocutionary force of an assertion.

Therefore, the speech acts by which subjects express a mental state (e.g., the utterance "p") and the speech acts by which subjects report that very same mental state (e.g., the utterance "I believe that p") have different truth-conditions insofar as they have different content (e.g., "p" and "I believe that p", respectively). Then, if expressions and reports of mental states have different content, and hence, different truth-conditions, why do they have similar assertion-conditions? Why can I say, for instance, "I believe that it is raining" or "I am grateful to you" (reporting so my mental state) in each conversational context in which I can say "It is raining" or "Thanks" (expressing so my mental state)?

According to Rosenthal, a mental state can be expressed both non-linguistically and linguistically. We just saw how a mental state can be linguistically expressed by a speech act (e.g., "It is raining", "Thanks" "To get wet would be inconvenient"). But mental states can be non-linguistically expressed as well. For instance, the belief that it's going to rain and the desire not to get wet can be non-linguistically expressed by picking up the umbrella before leaving the apartment. When a mental state is expressed non-linguistically, the mental state can be unconscious (i.e., I can pick up the umbrella *absent-mindedly* without thinking about the rain), but when a mental state is expressed linguistically, the mental state is almost always conscious[52]. Since to linguistically express a mental state involves consciousness, every subject with normal conceptual and linguistic abilities is supposed to be able to *non-inferentially report* a mental state (e.g., by asserting "I believe that p") in the same conversational contexts in which she is able to *linguistically express* that mental state (e.g., by asserting "p"). As a result, it becomes *second-nature* for us (for me and for the rest of the speakers of the language) that both the linguistic expressions of a mental state and the non-inferential reports of that mental state are uttered in the same conversational contexts or under the same conditions. Then, it becomes second-nature for us that the expression of belief "p" and the report of belief "I believe that p" have the same assertion-conditions.

Moore's paradox is explained by attending to this second-nature, embedded in the usages of the language, by which the assertion "p" is asserted under the same conditions as the non-inferential report "I believe that p". On the one hand, it is irrational to assert the Moore's sentence "p, but I don't believe that p", in spite of the fact that the truth-conditions of "p" have to do with the fact that p and the truth-conditions of "I don't believe that p" have to do with the fact that I don't have any belief about p, because to assert "p" and "I don't believe that p" in the same speech act violates their conditions of assertion. By asserting "p", I consciously express my belief that p, and so, as long as I have a normal linguistic dexterity, I am supposed to be able to appropriately and non-inferentially report that I believe that p. However, instead of asserting "I believe that p" (reporting so that I have the belief that p), I assert "I don't believe that p" (reporting so that I don't have any belief about p). Thus, to assert the Moore's sentence "p, but I don't believe that p" is irrational because it violates (at least in the conversational contexts in which Moore's paradox arises) the appropriate assertion-conditions of both "p" and "I don't believe that p". On the other hand, it is irrational to assert the Moore's sentence "p, but I believe that not-p", in spite of the fact that the truth-conditions of "p" have to do with the fact that p and the truth-conditions of "I believe that not-p" have to do with the fact that I have the belief that not-p, because to assert "p" and "I believe that not-p" in the same speech act violates their conditions of assertion. By asserting "p", I consciously express my belief that p, and so, as long as I have a normal linguistic dexterity, I am supposed to be able to appropriately and

[52] Rosenthal thinks that there is an exception to this fact: expressions of second-order beliefs. "I believe that p" expresses the second order belief "I believe that p". But, according to Rosenthal, when we say "I believe that p" we are normally conscious of the first-order belief "p" and not of the second-order belief "I believe that p" (which is the belief expressed by the assertion).

non-inferentially report that I believe that p. However, instead of asserting "I believe that p" (reporting so that I have the belief that p), I assert "I believe that not-p" (reporting so that I have the belief that not-p). Thus, to assert the Moore's sentence "p, but I believe that not-p" is irrational because it violates (at least in the conversational contexts in which Moore's paradox arises) the appropriate assertion-conditions of both "p" and "I believe that not-p".

Since to assert "p" linguistically expresses *my present* belief that p, to assert "p" involves consciousness of *my present* belief that p. That's why it is second-nature for us to assert "I believe that p", and not "I believed that p" or "She believes that p", whenever we would have asserted "p"; i.e., that's why "I believe that p", and not "I believed that p" or "She believes that p", has the same assertion-conditions as "p". Then, it is not irrational to assert or to judge the content of "p, but I didn't believe that p" and "p, but she doesn't believe that p" because it doesn't involve the violation of the assertion-conditions of any of its components.

Moreover, unlike Moore and Baldwin, Rosenthal promises to explain, not only why it is irrational to *assert* a Moore's sentence, but also 4) why it is irrational to *silently judge* its content in thought. To begin with, Rosenthal admits that pure pragmatic accounts don't seem to be able to explain why it is irrational to silently judge a Moore's sentence insofar as they rely on the assertion-conditions of Moore's sentences. Indeed, if it is second nature for us to assert "I believe that p" whenever we would have asserted "p", it is due to the fact that linguistic expression involves conscious expression. By contrast, Rosenthal admits, it is not clear that the "thought-conditions" of thinking "p" and thinking "I believe that p" are the same because we can think that p even unconsciously, that is, we can think "p" without thinking "I believe that p".

> "[…] the tie between speech acts and the intentional states they express cannot by itself explain why it should be impossible to think Moore's paradox. That's because the tie between speech acts and intentional states cannot explain anything about the mental analogue of assertibility conditions. Nor does that tie itself have any suitable mental analogue. Because asserting expresses a corresponding belief, it's impossible to assert anything without believing it. But it's plainly possible to think something without thinking that one thinks it. Indeed, I've urged that this typically happens when our thoughts aren't conscious. So we cannot explain the impossibility of thinking Moore's

> paradox by appeal to the same factors that explain the impossibility of saying it." (Rosenthal, 2005, p. 274).

To absent-mindedly pick up the umbrella when one is leaving the apartment is one of the examples that Rosenthal gives to explain unconscious thoughts or beliefs. One might be unconscious of the fact that one thinks that it is raining outside because one doesn't have the second-order thought or the second-order belief "I think/I believe that it is going to rain". However, the fact that one has the unconscious first-order thought that it is raining is supposed to explain why one picks up the umbrella (absent-mindedly and without realizing) before leaving the apartment. Therefore, the thought "p" and the thought "I believe that p" don't have the same "thoughts-conditions" (even though the assertion "p" and the assertion "I believe that p" have the same assertion-conditions) because the thought "p" can be unconscious; i.e., it can occur without occurring "I believe that p" as well.

At this point, it should be obvious that Rosenthal's account needs an additional claim to explain why it is irrational to judge the content of a Moore's sentence silently in thought. This additional claim has to do with *rationality*. According to Rosenthal, even if it is possible to think that p without thinking that one thinks that p, it is irrational to think that p while also thinking that one doesn't think that p or that one thinks that not-p. Then, to judge the content of the sentences "p, but I don't believe that p" and "p, but I believe that not-p" is irrational because, once the subject thinks "p", rationality prescribes *not* to think either "I don't think that p" or "I think that not-p".

> "We can perfectly well have the thought that *p* without thinking that we have it. But it's irrational to think both that *p* and that one doesn't think that *p*. So, if the question arises about whether one thinks that *p* and one does actually think it, it would then be irrational to hold that one doesn't. So it would be irrational to have an assertoric thought that conjoined those two contents. Only insofar as we are rational in this particular way is thinking Moore's paradox absurd." (Rosenthal, 2005, p. 276).

Is this account satisfactory? It is certainly better than Moran's and Baldwin's account insofar as it somehow explains the fourth desideratum of Moore's paradox (i.e., why it is irrational, not only to assert, but also to silently judge the content of a Moore's sentence). However, Rosenthal's explanation of the fourth desideratum is not completely satisfactory. Insofar as the kind of irrationality that takes place when asserting a Moore's sentence and when silently judging the content of a Moore's sentence seems to be the same kind of irrationality, it seems that Moore's paradox is the same phenomenon both when it arises at asserting a Moore's sentence and when it arises at silently judging the content of a Moore's sentence. Then, it seems that a good account of Moore's paradox should be able to provide a homogenous explanation of both the asserted and the thought paradox. However, Rosenthal's account explains the asserted paradox by attending to the *assertion-conditions* of the sentences "p" and "I believe that p", while it explains the thought paradox by attending to the *irrationality* of thinking "p" in the same act as thinking "I don't believe that p" or "I believe that not-p". Thus, Rosenthal's pragmatic account of Moore's paradox (i.e., based on assertion-conditions) seems to collapse into a psychological account of Moore's paradox (based on the subject's rationality and internal consistency) when dealing with the "thought" paradox. As a result, any account of Moore's paradox able to explain the phenomenon in a homogeneous way will have an advantage (at least in that regard) against Rosenthal's account of Moore's paradox.

Finally, Hamilton (2014) attributes to Wittgenstein a pragmatic view of Moore's paradox that seems to solve the problem of lack of homogeneity faced by Rosenthal's account. According to this view, Moore's paradox is an instance of a phenomenon labelled *pragmatic self-defeat*. While in cases of self-refutation subjects contradict or refute themselves (e.g., "It is raining and it is not raining"), in cases of pragmatic self-defeat subjects don't contradict themselves because they don't even make a judgement: they violate the conditions of sense, and so, their *utterances* or *thoughts* are nonsensical (neither true nor false). Pragmatic self-defeat is a first-person phenomenon because it arises in cases of facts about oneself that can be described by a third-person subject but that cannot be expressed or described by oneself. For instance, the utterance or thought "Maybe I don't understand the meaning of my own words" is an instance of pragmatic self-defeat because a *third-person* can appropriately describe the fact that I may not understand the meaning of my own words, but it is nonsensical *for me* to express or to describe the fact that I may not understand the meaning of the words that I am pronouncing because, by doing so, I would have shown that the opposite of what I express or describe is true. Then, from the idea of pragmatic self-defeat could be explained why the sentences "p, but I don't believe that p" and "p, but I believe that not-p" are irrational both to utter and to think: the fact that p is the case but I don't believe that p and the fact that p is the case but I believe that not-p can be appropriately described by a third-person, but they cannot be expressed or described by me because to utter or to think “p, but I don’t believe that p” or “p, but I believe that not-p” is nonsensical insofar as it violates the conditions of sense.

However, the problem with this account of Moore’s paradox is that Moore’s sentences don’t seem to be analogous to sentences like “Maybe I don’t understand the meaning of my own words”, and so, they don’t seem to be cases of pragmatic self-defeat. It is true that I cannot say or think about myself that I don’t understand the meaning of my current words without violating the conditions of sense because I cannot express or describe (what involves using words) the fact about me that I may not understand the meaning of the words that I am pronouncing or thinking, neither from the first-person deliberative perspective nor from the third-person self-inspective perspective. However, things are different in the case of “p, but I don’t believe that p” and “p, but I believe that not-p”. It is true that I cannot judge or assert those sentences from the first-person deliberative perspective without irrationality, but I actually can judge or assert those sentences without irrationality from the third-person self-inspective perspective to describe the fact that p is the case but I don’t believe that p, and the fact that p is the case but I believe that not-p. In fact, Tom’s utterance or thought “Men and women are equal, but I don’t believe so” is an example of a subject that, without violating the sense conditions of the utterance or thought, describes from the third-person perspective of self-inspection the fact that men and women are equal but he doesn’t have the belief that men and women are equal. (This idea will be explored in more detail in the following sections). As a result, claiming that Moore’s paradox is a case of pragmatic self-defeat leaves unexplained the desideratum 3), namely, that Moore’s sentences (which are in the first-person present tense) may be true$_{rs}$ sometimes.

Then, pragmatic accounts of Moore’s paradox are not able to explain the phenomenon in an appropriate way because they have problems explaining the fourth desideratum (i.e., that Moore’s paradox can arise both when asserting and when judging in thought the content of a Moore’s sentence) or the third desideratum (i.e., that Moore’s sentences may be true$_{rs}$). Moore’s and Baldwin’s pragmatic accounts don’t explain the fourth desideratum at all, Rosenthal’s account explains the fourth desideratum using two explanantia rather than one, and Wittgenstein’s account (in Hamilton’s view) explains it at the cost of wrongly assimilating Moore’s paradox to the phenomenon of pragmatic self-defeat, leaving so unexplained the third desideratum.

4.3 Psychological accounts of Moore's paradox

Psychological accounts are a cluster of accounts of Moore's paradox that have two ideas in common. Firstly, as all but semantic accounts, they think that the irrationality of Moore's paradox has nothing to do with the truth-conditions of Moore's sentences: they are not semantic contradictions, and so, they can be true$_{\mathrm{rs}}$ under certain circumstances. Secondly, they think that the irrationality of Moore's paradox arises (ultimately[53]) because to assert or to judge the content of a Moore's sentence from a first-person perspective involves a lack of psychological consistency in the subject. Depending on the kind of psychological inconsistency thought to be responsible for the irrationality of Moore's paradox, there are two different kinds of psychological accounts. On the one hand, some psychological accounts of Moore's paradox claim that the inconsistency responsible for the irrationality takes place among the *conscious* mental states of the subject. On the other hand, other psychological accounts of Moore's paradox claim that the inconsistency responsible for the irrationality takes place among the *commitments* endorsed by the subject.

4.3.1 Psychological accounts based on consciousness

Some psychological accounts (Baldwin, 1990[54]; Williams, 2006; Shoemaker, 1996) explain Moore's paradox by using consciousness. They claim that the irrationality of Moore's paradox arises because asserting or judging the content of a Moore's sentence from the first-person perspective involves having inconsistent or contradictory *conscious beliefs*, where consciousness is understood as having a second-order belief about a first-order mental state

[53] For some defenders of this kind of psychological accounts (e.g., Williams, 2006) think that, even if the irrationality of Moore's paradox arises ultimately because of the lack of psychological consistency of the subject in *judging* the content of a Moore's sentence, the irrationality of *asserting* a Moore's sentence requires a different explanation dependent of the irrationality of judging its content.

[54] In the same text as Baldwin offers his pragmatic account of Moore's paradox, he offers a psychological account of Moore's paradox based on consciousness that he attributes to Wittgenstein.

(e.g., I am conscious of my first-order belief "It is raining" if I have the second-order belief "I believe that it is raining"). Shoemaker is among the defenders of a psychological account of Moore's paradox based on the idea of consciousness and his account is going to be discussed henceforth as a paradigm of psychological accounts based on consciousness.

Shoemaker explains Moore's paradox by using the concept of *mental assent*, which is considered to be the mental correlate of linguistic assertions (when they are sincere). According to Shoemaker, the assent-conditions of "p" are similar to the assent-conditions of "I believe that p" because assenting to "p" involves having *available* both the belief that p and the belief that I believe that p (i.e., the second-order belief), and so, it involves consciousness. The justification for this idea, according to Shoemaker, is the *self-intimation* claim: that there is a *constitutive relation* between having available the belief that p and believing that one believes that p (i.e., being conscious of one's first-order belief) because *rational subjects* who have available the belief that p will be disposed to use the propositional content "p" both in their theoretical and practical reasoning, and so, they will be disposed to act as it corresponds to the belief that p, which includes being disposed to utter the first-person avowal "I believe that p" (i.e., a self-ascription of the second-order belief that p). For instance, if I have available the first-order belief that it is raining, I have the disposition to pick up the umbrella because I don't want to get wet and the disposition to say "I believe that it is raining" (second-order belief) to tell others why I'm picking up the umbrella, among other practical and theoretical dispositions.

Indeed, Shoemaker endorses a functionalist view of the nature of mental states according to which mental states are "core realizations" defined by their causal relations and typically implemented in neurobiological states. For instance, the mental state of pain is considered to be a neurobiological state (i.e., firing C-fibers) that is typically caused by tissue damage (input) and that typically causes a moan or a flinch (output). In the case of rational animals (i.e., human beings), having available a first-order mental state always involves consciousness or second-order belief (i.e., self-intimation claim). This consciousness or second-order belief, however, doesn't involve any *causal relation* between the first-order mental state and the second-order belief; by contrast, the second-order belief or consciousness is an *aspect* of the core realization of the first-order mental state itself (i.e., constitutive claim) when that first-order mental state is available to a rational animal (i.e., human beings). Thus, this consciousness or second-order belief can be causally determined in the core realization of the first-order mental state itself because it has *sui generis* causal effects (outputs). For instance, the mental state of pain causes the desire to stop having pain, and so, it causes avoidance of the source of pain; but it also causes in rational animals that the subject takes an aspirin or that she utters the first-person avowal "I have a terrible headache". The latter two, but not the former, are considered causal effects of the second-order belief or consciousness included in the core realization of the first-order mental state of pain when it is available to a rational subject.

Then, Moore's paradox is explained as follows. Assenting to "p" involves having available the belief that p and being conscious (i.e., having a second-order belief) of the fact that one believes that p. So, on the one hand, to judge the content of "p, but I don't believe that p" is irrational because it would involve having *contradictory conscious* beliefs (what, according to Shoemaker, is an impossible psychological state). For assenting to "p" involves having available the belief that p and having the second-order belief that one believes that p, and this second-order belief (i.e., the belief that I believe that p) is contradictory with the second-order belief available at assenting to "I don't believe that p" (i.e., the belief that I don´t believe that p). So, assenting to "p, but I don't believe that p" would involve *consciously* having two contradictory beliefs: the second-order beliefs "I believe that p" and "I don't believe that p". On the other hand, to judge the content "p, but I believe that not-p" is irrational because it involves having *inconsistent conscious* beliefs (which, according to Shoemaker, is a possible but irrational psychological state). For asserting to "p" involves having available the belief that p and having the second-order belief that one believes that p, and this second-order belief (i.e., the belief that I believe that p) is inconsistent with the belief available at assenting to "I believe that not-p" (i.e., the belief that I believe that not-p). So, assenting to "p, but I believe that not-p" involves *consciously* having two inconsistent beliefs: the second-order beliefs "I believe that p" and "I believe that not-p". Moreover, since to judge the content of a Moore's sentence involves contradictory or inconsistent beliefs and what cannot be coherently believed cannot be coherently asserted (Shoemaker, p. 76), it is explained as well why Moore's sentences are irrational to *assert*. Therefore, unlike some pragmatic accounts, Shoemaker's psychological account explains why it is irrational both to assert and to judge the content of a Moore's sentence in a homogeneous way (i.e., fourth desideratum).

Also, insofar as *presently* assenting to "p" doesn't involve having available either the belief that I believed that p or the belief that she believes that p, assenting to "p, but I didn't believe that p" or "p, but she doesn't believe that p" is not irrational because there are no contradictory or inconsistent conscious beliefs involved.

Psychological accounts of Moore's paradox based on consciousness may seem plausible. However, I will argue that they are too broad to account for the first desideratum of Moore's paradox in an appropriate way (i.e., why it is sometimes irrational to assent to a Moore's sentence) because they mistakenly predict that it is irrational to assent to a Moore's sentence in some cases in which Moore's sentences are not irrational to assent to. Remember the case of Tom, a man who is self-deceived about the fact that men and women are equal. Tom judges "Men and women are equal" when he deliberates from the first-person perspective about gender equality (for he doesn't find real reasons to judge otherwise). However, by self-inspecting himself from the third-person perspective, Tom finds certain attitudes and behaviours pointing to the fact that he doesn't actually believe that men and women are equal. For instance, he realizes that he expects women to share more homemade food than men when doing a picnic, that he tends to rely on men rather than women when he needs help for an intellectual task or that he expects women to do most of the domestic chores. Thus, Tom concludes his self-inspection judging "I don't believe that men and women are equal". In Tom's situation, assenting to the Moore's sentence "Men and women are equal, but I don't believe so" is not an instance of Moore's paradox because such a sentence is not irrational to assent to (i.e., *the act* of assenting to that Moore's sentence is not irrational in this case). It is true that Tom is in an irrational cognitive state insofar as he *judges* that p is the case when deliberating about whether p but he is unable to *believe* so[55] (as he finds out by the third-person process of self-inspection); however, even if Tom is in an irrational cognitive state, it is not irrational for him to assent to the Moore's sentence "p, but I don't believe that p" in this situation because "I don't believe that p" is the result of self-inspection. I will argue that the problem with psychological accounts of Moore's paradox based on consciousness is that they wrongly predict that Tom's assent to "Men and women are equal, but I don't believe so" must be an instance of Moore's paradox when it is clear that it is not.

Let's see in detail how psychological accounts based on consciousness wrongly predict that Tom's assertion "Men and women are equal, but I don't believe so" is an instance of Moore's paradox. On the one hand, when assenting to "Men and women are equal" (the first part of the sentence), Tom must have the available first-order belief that men and women are equal and the second-order belief that he believes that men and women are equal. Secondly, when assenting to "I don't believe that men and women are equal" (the second part of the sentence), Tom must have available the second-order belief that he doesn't believe that men

[55] Actually, Tom has a self-deceived belief.

and women are equal. As a result, Tom must have the second-order beliefs "I believe that men and women are equal" and "I don't believe that men and women are equal"; i.e., Tom must *consciously* have two contradictory beliefs (what, according to Shoemaker, is an impossible psychological state). It follows that Tom's assent to "Men and women are equal, but I don't believe that they are" must be irrational to make, and hence, that it must be an instance of Moore's paradox. Thus, psychological accounts of Moore's paradox based on consciousness are too broad to appropriately explain the first-desideratum of Moore's paradox (i.e., why it is irrational to assent to a Moore's sentence): it follows from their *explanans* (i.e., that Moore's paradox has to do with contradictory or inconsistent conscious beliefs) that some cases in which it is not irrational to assent to a Moore's sentence must be cases in which it is irrational to assent to a Moore's sentence.

Two replies are open to the defender of a psychological account based on consciousness to counteract Tom's counter-example. The first reply is that the objection from the last paragraph is faulty because psychological accounts based on consciousness claim that second-order beliefs grant consciousness about a first-order mental state or about the lack of a first-order mental state *when and only when* the second-order belief has been formed from the first-person perspective (i.e., by deliberating about whether p) and not from the third-person perspective (i.e., by a process of self-inspection). What happens in Tom's case, the reply continues, is that Tom has the unconscious belief that men are superior to women (unconscious belief that allegedly explains his non-egalitarian behaviour). And, since Tom forms the second-order belief that he doesn't believe that men and women are equal from the third-person perspective (i.e., from self-inspection), this second-order belief is not supposed to grant consciousness to any first-order belief or lack of first-order belief. Therefore, the reply concludes, psychological accounts of Moore's paradox based on consciousness don't predict that it is irrational to assent to "Men and women are equal, but I don't believe so" in Tom's case because Tom doesn't have contradictory or inconsistent conscious beliefs: he has the second-order beliefs "I *believe consciously* that men and women *are* equal" and "I *don't believe consciously* that men and women *are not* equal [because I believe it only unconsciously]". Then, it is not the case that Tom has the contradictory second-order beliefs "I believe consciously that men and women are equal" and "I don't believe consciously that men and women are equal".

However, this reply doesn't work as it stands because it is based on a claim that has not been justified by psychological accounts of Moore's paradox based on consciousness.

Psychological accounts based on consciousness need to explain why second-order beliefs are supposed to grant consciousness about a first-order mental state or about the lack of a first-order mental state *when and only when* second-order beliefs are formed from a first-person deliberative perspective and not when they are formed from a third-person self-inspective perspective. Namely, psychological accounts based on consciousness need to explain why Tom's assent to the second part of the sentence (i.e., "I don't believe that men and women are equal") made from the third-person perspective (i.e., by self-inspecting himself) is supposed to involve *the second-order belief* "I unconsciously believe that men and women are not equal" *without involving consciousness* of the first-order belief "Men and women are not equal". Indeed, psychological accounts based on consciousness explain consciousness by the occurrence of a second-order belief and Moore's paradox by the occurrence of inconsistent or contradictory conscious beliefs. If it happens that there are cases in which second-order beliefs don't grant consciousness because they are formed by the third-person process of self-inspection and not by first-person deliberation, psychological accounts based on consciousness have to tell us which other elements characteristic of the first-person perspective need to be added to a second-order belief to deliver consciousness. Once they find out which elements characteristic of the first-person deliberative perspective are necessary, together with the second-order belief, to deliver consciousness of a first-order mental state, they should incorporate those new elements into their explanans of Moore's paradox. Meanwhile, their accounts are incomplete, and what is incomplete neither it is fully explanatory nor can be appropriately refuted.

The second reply to Tom-like counter-examples is offered by Shoemaker himself, who was aware of the existence of Tom-like cases[56]. According to Shoemaker, what happens in cases like Tom's is that the subject suffers from a failure of unity of consciousness so that his mind is somehow divided into two parts, a *speaking part* and a *non-speaking part*, and each of these parts hosts a different *subject of mental states* (Shoemaker, 1996, p. 90). Then, if Tom's assent to the Moore's sentence "Men and women are equal, but I don't believe so" is not irrational (i.e., if it is not an instance of Moore's paradox), it is because the speaking part of Tom's mind has the conscious belief that men and women are equal (that's why he avows:

[56] […] suppose that a psychiatrist tells me that I have the repressed belief that I was adopted as an infant. In fact, the psychiatrist has confused me with another patient (he has been reading the wrong case history), and has no good grounds for this belief attribution. But I accept it on his authority. It seems compatible with this that when I consider the proposition I am supposed to believe, that I was adopted, I find no evidence in its support, and am disposed to deny it. Here, it seems, I might be in a position to assert "I believe that I was adopted, but that's not true." (Shoemaker, 1996, p. 89)

"Men and women are equal" from the deliberative first-person perspective) and, at the same time, the speaking part of Tom's mind claims that the non-speaking part doesn't have the belief that men and women are equal (so that the speaking part says "I don't believe that they are" referring to the non-speaking part) because the non-speaking part has the unconscious belief that men and women are not equal. This is supposed to explain Tom's sexist behaviour as well. Thus, the reply concludes, Tom's case is not an instance of Moore's paradox because it is not a case in which a *single* subject assents both to "p" and to "I don't believe that p"; instead, there are two subjects or two subjectivities involved. However, this explanation of Tom's case doesn't seem plausible at all. There are independent reasons to question the possibility of the division of the subject's mind required by Shoemaker's reply (remember the discussion about the static and dynamic puzzles in section 3.2.1.) and Shoemaker himself doesn't offer any independent reason to justify that such division of the mind is possible or plausible (for the only reason offered by Shoemaker to accept the idea of the division of the mind is that it explains Tom-like cases). Thus, talking about two subjects or subjectivities in one subject to explain Tom-like counter-examples seems like an *ad hoc* strategy with more costs in terms of parsimony and plausibility than explicative benefits.

Therefore, psychological accounts of Moore's paradox based on consciousness cannot appropriately explain the first desideratum of a good account of Moore's paradox: why it is sometimes irrational to assent to a Moore's sentence. For it wrongly predicts that Tom's assent to "Men and women are equal, but I don't believe so" must be an instance of Moore's paradox (insofar as it must be irrational to assent) when it clearly isn't an instance of Moore's paradox (insofar as it is not irrational to assent).

4.3.2 Psychological accounts based on commitments

Psychological accounts based on commitments claim that assenting to a Moore's sentence from the first-person perspective is irrational because it involves the subject's endorsement of *inconsistent* (Baldwin, 2007) or *impossible* (Coliva, 2016) commitments. Baldwin claims that believing that p consists in being committed to the truth of "p". So, when a subject assents to "p", she forms the belief that p because she commits herself to the truth of "p". According to Baldwin, Moore's paradox is explained because assenting to a Moore's sentence involves endorsing inconsistent commitments, and so, inconsistent beliefs. On the one hand, assenting to "p, but I don't believe that p" is irrational because with "p" the subject commits herself to the truth of "p" at the same time that she says that she doesn't have any commitment about p (i.e., "I don't believe that p"). Thus, assenting to "p, but I don't believe that p" involves inconsistent commitments because it involves the commitment to the truth of "p" and the denial of that commitment: the commitment to the fact that one is not committed to the truth of "p". On the other hand, assenting to "p, but I believe that not-p" is irrational because with "p" the subject commits herself to the truth of "p" at the same time that she says that she has a commitment to the truth of "not-p" (i.e., "I believe that not-p"). So, assenting to "p, but I believe that not-p" involves inconsistent commitments because it involves the commitment to the truth of "p" and the commitment to the fact that one is committed to the truth of "not-p".

These commitments are considered by Baldwin to be "obviously inconsistent" and, insofar as they are endorsed by a subject in the same judgement or assertion, *irrational* to endorse. However, even if they are supposed to be irrational to endorse by a subject in the same act of assent, Baldwin considers that it is *possible* for a subject to have such commitments (so that they are not self-defeating or impossible to endorse)[57]. Moreover, since the propositional content of "p" is about the fact that p and the propositional content of "I don't believe that p" and "I believe that not-p" is about my lack of belief about p or about my belief that not-p, to assent to a Moore's sentence involves inconsistent commitments but not a semantic contradiction.

Baldwin's account seems to face two different problems to appropriately explain the first desideratum of Moore's paradox (i.e., why it is sometimes irrational to assent to a Moore's sentence). Firstly, it is not clear that Baldwin's account manages to explain the irrationality of assenting to a Moore sentence *at all*. Baldwin thinks that assenting to a Moore's sentence is irrational because it involves endorsing "obviously inconsistent" (but possible) commitments in the same act of assent. However, insofar as the commitments involved in assenting to a

[57] "We do of course find ourselves from time to time with inconsistent commitments, but it is absurd to make commitments whose inconsistency is obvious in the very judgement itself." (Baldwin, 2007, p. 86) By "absurd" here Baldwin has to mean "irrational" and not "contradictory". For the commitments involved in the assent of a Moore's sentence are considered to be commitments to different propositional contents (i.e., a commitment to p and a commitment to the fact that I am committed to not-p or that I have no commitments in regard to p, depending on the Moore's sentence).

Moore’s sentence are commitments to different semantic contents (i.e., a commitment to the truth of “p” and a commitment to the fact that I have no commitments to the truth of “p” or to the fact that I have a commitment to the truth of “not-p”, depending on the paradox), it is not obvious why those commitments are supposed to be inconsistent. Thus, insofar as it is not explained why the commitments involved in assenting to a Moore’s sentence are “obviously inconsistent” in spite of being commitments to different semantic contents, the irrationality characteristic of Moore’s paradox hasn’t been explained at all. Rather, it seems that what has been offered so far is a reformulation of the paradox (i.e., why it is sometimes irrational to assent to a Moore’s sentence if it is not supposed to be a self-contradiction) in terms of commitments (i.e., why it is sometimes irrational to endorse the commitments involved in assenting to a Moore’s sentence if they are not supposed to be self-contradictory—for they are possible to endorse and they are commitments to different semantic contents—). This reformulation of the paradox might be a step ahead towards its resolution, but it cannot be considered a complete account of Moore’s paradox until it is explained why the commitments considered to be involved in Moore’s paradox are inconsistent and irrational to hold in the same act of assent.

Secondly, it seems that Baldwin’s account cannot explain why Tom can assent to “Men and women are equal, but I don’t believe so” without giving rise to an instance of Moore’s paradox. On the one hand, since Tom finds evidence supporting the fact that men and women are equal, he concludes his deliberation about gender equality judging “Men and women are equal”. On the other hand, since Tom finds behavioural evidence against the fact that he believes that men and women are equal, he concludes his process of self-inspection judging “I don’t believe that men and women are equal”. As a result, Tom can assent to the Moore’s sentence “Men and women are equal, but I don’t believe so”, and so, he can commit himself both to the fact “Men and women are equal” and to the fact “I am not committed to the truth of men and women are equal” in the same act of assent. Thus, Baldwin’s account predicts that Tom’s assent to “Men and women are equal, but I don’t believe so” must be irrational to make (i.e., must be an instance of Moore’s paradox) because it involves, in a single act, what are considered to be inconsistent commitments. However, Tom’s assent to “Men and women are equal, but I don’t believe so” is not irrational to make, and so, it is not an instance of Moore’s paradox.

Coliva (2016) develops a psychological account of Moore’s paradox based on commitments that seems to solve the problems faced by Baldwin’s account. Instead of arguing (as Baldwin does) that Moore's paradox involves the subject's endorsement of inconsistent commitments (i.e., commitments which are irrational but *possible* to endorse by a subject—e.g., Tom's case—), Coliva claims that Moore's paradox involves the subject's endorsement of *impossible* or *self-defeating* commitments so that to endorse one of the commitments logically involves the destruction of the other. According to Coliva, subjects can have two kinds of beliefs: *beliefs as dispositions* and *beliefs as commitments*. A subject has the belief that p as disposition when she *behaves* as if she believed that p, even though "p" is not the conclusion of her first-person deliberation about whether p (either because she hasn't ever deliberated about whether p or because she has deliberated about whether p only to conclude something different from p). By contrast, a subject has the belief that p as commitment when p is the conclusion of her first-person deliberation about whether p so that the subject commits herself to the truth of "p"; i.e., she sees herself as *being bound* to the actions prescribed by the belief that p (regardless of whether she actually performs those actions or not).

Coliva's account of Moore's paradox differs from Baldwin's in that Coliva thinks that the commitments endorsed in assenting to a Moore's sentence are self-defeating and that both Moore's paradox and Tom-like cases can be accounted for by attending to the distinction between beliefs as dispositions and beliefs as commitments. Moore's paradox arises when subjects assent to a Moore's sentence from the *first-person deliberative perspective* (the realm of beliefs as commitments) because to do so would involve the subject's endorsement of impossible or self-defeating commitments. On the one hand, assenting to "p, but I believe that not-p" is irrational because with "p" the subject commits herself to the truth of "p" and with "I believe that not-p" the subject *undoes* that commitment insofar as she *self-ascribes* to herself the commitment to "not-p" (i.e., the subject commits herself to the fact that she is committed to not-p), a commitment that is considered to be logically incompatible with the commitment to "p". On the other hand, assenting to "p, but I don't believe that p" is irrational because with "p" the subject commits herself to the truth of "p" and with "I don't believe that p" the subject *undoes* that commitment insofar as she *self-ascribes* to herself the commitment to "open-mindedness" (lack of belief) about p (i.e., the subject commits herself to the fact that he has no commitments about whether p), a commitment that is considered to be logically incompatible with the commitment to "p". Therefore, assenting to a Moore's sentence is irrational (in spite of the fact that Moore's sentences are not semantic contradictions) because it would involve the subject's endorsement of impossible or self-defeating commitments: one commitment via self-ascription (i.e., "I believe that not-p" or "I don't believe that p") and other commitment via judgement or assertion about a fact of the world (i.e., "p").

Coliva strives to argue that assenting to a Moore's sentence from the first-person deliberative perspective would involve *logically impossible* or *logically self-defeating* commitments, as opposed to mere inconsistent commitments that subjects may *possibly* endorse (as Baldwin claims). Notice that if the commitments involved in Moore's paradox are impossible or self-defeating to endorse, it is explained why Moore's sentences are irrational to assent from the first-person deliberative perspective in a way that goes beyond the mere reformulation of the paradox. For commitments that are impossible or self-defeating to endorse by a rational subject cannot be anything but irrational commitments. To explain why the commitments involved in assenting to a Moore's sentence from the first-person deliberative perspective are impossible or self-defeating to endorse, Coliva argues that having the belief as commitment that p entails *seeing oneself as being bound* to the kind of actions (including both speech acts and actions) that are mandated by the truth of "p", regardless of whether the subject actually implements those particular actions or not:

> "[…] having a belief as a commitment consists in knowingly and willingly binding oneself to those courses of action that 'are entailed by those desires and beliefs by the light of certain normative principles of inference'. If one does not comply with them, one will be held responsible for not doing so and will have to be self-critical or accept criticism from others for it. Thus, to have a belief as a commitment entails seeing oneself as having to implement a certain behaviour (and accepting criticism for not 'living up to one's commitments' should one fail to behave accordingly)." (Coliva, 2016, p. 258).

Thus, it is logically impossible for a subject to have the belief as commitment that p and to sincerely assent to "not-p" or to "open-mindedness" about p (from the deliberative first-person perspective, the realm of commitments). For, insofar as she has the belief as commitment that p, she has to see herself as mandated by the courses of action prescribed by the truth of "p", and so, if she sincerely assents to "not-p" or to "open-mindedness" about p from the deliberative first-person perspective, she has to see herself as automatically undoing her commitment to the truth of "p". Since beliefs as commitments, unlike beliefs as dispositions, have to do only with seeing oneself as being bound to certain actions and not with actually performing those actions, a subject who has the belief as a commitment that p can *knowingly and willingly* perform actions against what is mandated by that belief, without dropping her belief as commitment because of that, under the condition that she feels compelled to act as it is prescribed by the truth of "p" and that she accepts criticism for acting against what is prescribed by the truth of "p" (Coliva, 2016, pp. 258-259).

In regard to Tom-like cases, Coliva tries to explain them with the distinction between beliefs as commitments and beliefs as dispositions[58]. On the one hand, Tom has the *belief as disposition* that men and women are not equal because he *acts* in the way that is characteristic of having the belief as disposition that men and women are not equal (e.g., expecting women to do the chores, preferring men for intellectual tasks, etc). On the other hand, Tom has the belief as commitment that men and women are equal because he *judges* that men and women are equal when deliberating from the first-person perspective about gender equality, and so, he sees himself as being bound to act as it is prescribed by the belief that men and women are equal and he accepts criticism when he acts against that belief. As a result, Tom has both the belief as disposition that men and women are not equal and the belief as commitment that men and women are equal, and so, he can assent to "Men and women are equal, but I believe that men and women are not equal" (i.e., "p, but I believe that not-p") without giving rise to an instance of Moore's paradox. When Tom assents to "Men and women are equal", he commits himself to the fact that men and women are equal. And when Tom assents to "I believe that men and women are not equal", Tom self-ascribes to himself the unconscious belief as disposition that men and women are not equal, committing so himself to the fact that he has the unconscious belief as disposition that men and women are not equal. Since the first commitment is to the truth of "Men and women are equal" and the second commitment is to the truth of "I have the unconscious belief as disposition that men and women are not equal", no impossible or self-defeating commitments are involved here. Thus, it is explained why Tom can assent to "Men and women are equal, but I believe that men and women are not equal"

[58] Coliva puts the example of Jane's case, which is similar to Tom's case:

> "The contrast between beliefs as commitments and dispositions, then, could be illustrated by Jane's situation. She finds herself with a belief as a disposition that her husband is unfaithful to her. That disposition shapes much of her behaviour and can have various causes. However, she also has and avows her belief as a commitment, held on the basis of evidence that she herself has assessed, that he is not. The latter belief exerts normative force on her and, consequently, she ought to try to get rid of her recognisably irrational disposition. In the end, he might not be able to overcome it (completely)." (Coliva, 2016, p. 258)

without giving rise to an instance of Moore's paradox (i.e., without being such assent irrational to make).

However, Coliva's account has problems explaining the other version of Tom's case: how it is that Tom can assent to "Men and women are equal, but I don't believe so" without giving rise to an instance of Moore's paradox (i.e., without being such assent irrational to make). Indeed, insofar as a third-person process of self-inspection can conclude both with a self-ascription of belief (e.g., "I believe that p" or "I believe that not-p") and with a self-ascription of lack of belief or "open-mindedness" (e.g., "I don't believe that p " or "I don't believe that not-p"), Coliva has to admit that there are two versions of Tom-like cases that correspond to the two versions of Moore's paradox: a version in which the subject assents to "p, but I believe that not-p" without giving rise to an instance of Moore's paradox and a version in which the subject assents to "p, but I don't believe that p" without giving rise to an instance of Moore's paradox. We have seen that Coliva's account seems to explain appropriately the version of Tom's case in which Tom assents to "p, but I believe that not-p" without giving rise to an instance of Moore's paradox. However, insofar as in Coliva's account the concept of belief as commitment is characterized as having to do only with *seeing oneself as being bound* to the kind of actions prescribed by the belief that p (regardless of how the subject actually acts), it seems that Coliva's account cannot explain appropriately why Tom can assent to "Men and women are equal, but I don't believe so" (i.e., "p, but I don't believe that p") without giving rise to an instance of Moore's paradox.

Let's see what follows from Coliva's account about this version of Tom's case. On the one hand, Tom assents to "Men and women are equal" because he forms the *belief as commitment* that men and women are equal when he deliberates from the first-person perspective about gender equality. In doing that, Tom is supposed to commit himself to the truth of "Men and women are equal", and so, he is supposed to see himself as being bound to the courses of action prescribed by the belief that men and women are equal and to accept criticism when he acts against them. On the other hand, Tom assents to "I don't believe that men and women are equal" because, when he self-inspects himself on the basis of his own behaviour (e.g., expecting women to do the chores, preferring men for intellectual tasks, etc), he concludes with a self-ascription of lack of *belief as commitment* about gender equality. Notice that Tom's assent to "I don't believe that men and women are equal" cannot be understood in this case as a self-ascription of lack of *belief as disposition* because Tom judges by self-inspection that he *acts* as if he believed that men and women are not equal, and so, according to Coliva's account, Tom must have the belief as disposition that men and women are not equal. Thus, Tom's assent to "I don't believe that men and women are equal" must be understood in this case as a commitment to the fact that one doesn't have any *commitment* about whether men and women are equal (i.e., as a self-ascription of lack of belief as commitment) rather than as a commitment to the fact that one doesn't *act* against the way that is characteristic of the belief as disposition that men and women are not equal (i.e., as a self-ascription of lack of belief as disposition). As a result, Coliva's account wrongly predicts that Tom's assent to "Men and women are equal, but I don't believe so" must be an instance of Moore's paradox (i.e., must be irrational to make) because, at assenting so, Tom is supposed to endorse the two self-defeating or impossible commitments that are considered to be characteristic of Moore's paradox: the commitment to the fact that men and women are equal and the commitment to the fact that one doesn't have any commitment about whether men and women are equal.

Thus, the problem with Coliva's account is that the distinction between belief as commitments and beliefs as dispositions doesn't seem to give appropriate account of the distinction between the first-person and the third-person perspectives, as it is proved by the fact that that distinction cannot appropriately explain the version of Tom-like examples in which the subject assents to "p, but I don't believe that p" without giving rise to an instance of Moore's paradox. Indeed, in order to explain the version of Tom-like cases in which the subject assents to "p, but I don't believe that p", the distinction between belief as commitments (i.e., seeing oneself as being bound to the courses of action prescribed by a belief) and beliefs as disposition (i.e., actually performing the actions prescribed by a belief) needs to be dropped. For this distinction detaches the fact of having the belief that p from the fact of actually performing the actions characteristic of the belief that p, and this detachment is incompatible with the appropriate account of why Tom can assent to "Men and women are equal, but I don't believe so" without giving rise to an instance of Moore's paradox.

This version of Tom-like cases is explained as follows. On the one hand, Tom assents to "Men and women are equal" because he concludes judging "Men and women are equal" when he deliberates from the first-person perspective about gender equality on the basis of evidence about whether men and women are equal or not. On the other hand, Tom assents to "I don't believe so" because, when he self-inspects himself from the third-person perspective on the basis of his own behaviour, he concludes his self-inspection judging that he doesn't have the *belief* that men and women are equal, in spite of the fact that he *judges* that men and women are equal, because he doesn't *act* in a compatible way with having the *belief* that men and women are equal. As a result, if one doesn't detach the notion of belief from the actions that are actually performed by a subject (as Coliva does with the distinction between belief as commitment and belief as disposition), it is possible to explain why Tom can assent to "Men and women are equal, but I don't believe so" without giving rise to an instance of Moore's paradox (i.e., without being irrational to assent so): Tom assents to "Men and women are equal" because he judges that men and women are equal when deliberating about gender equality and Tom also assents to "I don't believe so" because he judges by self-inspection that he doesn't qualify as believing that men and women are equal, even if he sincerely judges so, because he doesn't act in a compatible way with having the belief that men and women are equal.

Therefore, Coliva's account fails to explain the first desideratum of Moore's paradox (i.e., that sometimes it is irrational to assent to a Moore's sentence) combined with the second desideratum (i.e., that there is a conceptual difference between the two types of Moore's sentences) because Coliva's account wrongly predicts that Tom's assent to "p, but I don't believe that p" must be an instance of Moore's paradox (i.e., must be irrational to make) when it is not. It is true that Coliva's account appropriately predicts that Tom's assent to "p, but I believe that not-p" is not an instance of Moore's paradox (i.e., because it is not irrational to make), but to appropriately explain the second desideratum of Moore's paradox (i.e., that there is a conceptual difference between the two types of Moore's sentences) is necessary to predict appropriately when it is irrational to assent and when it is not irrational to assent to *both* "p, but I don't believe that p" and "p, but I believe that not-p".

Then, all psychological accounts of Moore's paradox (regardless of whether they are based on consciousness or on commitments) have in common that they are unable to appropriately explain the first desideratum of Moore's paradox (i.e., that sometimes assenting to a Moore's sentence is irrational). Shoemaker's account fails to explain the first-desideratum because it mistakenly predicts that Tom-like cases must be instances of Moore's paradox; Baldwin's account doesn't explain the first desideratum both because it is not clear that it explains the irrationality of Moore's paradox rather than giving a new formulation of the paradox in terms of commitments and because it doesn't explain Tom-like cases; and Coliva's account doesn't explain the first desideratum, when it is understood in the context of the second-desideratum (i.e., that there is a conceptual difference between the two Moore's sentences), because it predicts that the version of Tom-like examples in which the subject assents to "p, but I don't believe that p" are instances of Moore's paradox when they are not.

4.4 Epistemic accounts of Moore's paradox

Epistemic accounts of Moore's paradox (Fernández, 2005, 2013; Moran, 1997, 2001) claim that assenting to a Moore's sentence from the first-person deliberative perspective is irrational because it somehow involves a failure in the Transparency procedure responsible for first-person epistemic self-knowledge. As it was explained in the first chapter, Fernández claims that the bypass procedure is the first-person procedure of belief-formation responsible for Transparency and for first-person epistemic self-knowledge. As a reminder, a subject follows the bypass procedure, acquiring first-person self-knowledge, when she forms second-order beliefs in a transparent way: forming the second-order belief that she believes that p on the basis of the very same grounds on which she has formed the first-order belief that p. For example, I perform the bypass procedure, acquiring first-person self-knowledge, if I form my second-order belief that I believe that it is raining on the basis of the very same perceptual appearance of rain on which I have formed my first-order belief that it is raining. The details of how the bypass procedure is supposed to deliver (normally) first-person epistemic self-knowledge (i.e., strongly warranted true second-order beliefs) were explained in chapter one. What is relevant to us here is how Moore's paradox is supposed to be explained attending to the bypass procedure.

According to Fernández, assenting to a Moore's sentence from the first-person perspective is irrational, in spite of the fact that Moore's sentences can be true, because to do so involves an epistemic failure in the bypass procedure of belief-formation that the subject *must have been able to avoid*. On the one hand, assenting to "p, but I don't believe that p" is irrational because the subject is supposed to have formed the *first-order belief that p* on the basis of grounds that she takes to be appropriate to form that first-order belief that p, and so, according to the bypass procedure, she should have formed the *second-order belief that she believes that p* on the basis of those very same grounds. However, due to an epistemic failure in the bypass procedure responsible for first-person self-knowledge, the subject ends up forming the false *second-order belief that she doesn't believe that p*, as if she didn't have what she takes as appropriate grounds either for the first-order belief that p or for the first-order belief that not-p. On the other hand, assenting to "p, but I believe that not-p" is irrational because the subject is supposed to have formed the *first-order belief that p* on the basis of grounds that she takes to be appropriate to form that first-order belief that p, and so, according to the bypass procedure, she should have formed the *second-order belief that she believes that p* on the basis of those very same grounds. However, due to an epistemic failure in the bypass procedure responsible for first-person self-knowledge, she forms the *false second-order belief that she believes that not-p*, as if she had what she takes as appropriate grounds for the first-order belief that not-p. Thus, in both cases, the subject makes an epistemic error in the bypass procedure responsible for first-person self-knowledge that should have been obvious for her and that she should have been able to avoid. As a result, the subject is supposed to end up with a false second-order belief in both cases, and so, she is supposed to suffer from lack of first-person self-knowledge. Since the bypass procedure is applied only when forming beliefs about oneself in the present and from the first-person perspective, it is explained as well why it is not irrational to assent to "p, but I believed that not-p" or "p, but she believes that not-p".

However, insofar as Fernández's account explains Moore's paradox by an epistemic failure, the following conceptual problem seems to arise: that it doesn't seem possible to describe an example of a subject *who irrationally assents to a Moore's sentence from the first-person perspective* because of the occurrence of an epistemic failure in the bypass procedure of belief-formation. Indeed, if Moore's paradox were the result of an epistemic failure in the bypass procedure, that epistemic failure must be conceptually possible and it should be possible to describe an example in which a subject *irrationally assents to a Moore's sentence from the first-person perspective* as a result of such a mistake in the bypass procedure (for Moore's paradox arises only when the *act* of assenting to a Moore's sentence is irrational). However, neither does Fernández provide that example nor does it seem that we are able to find it.

Let's see this point comparing the case of Moore's paradox and the case of perception. According to Fernández, beliefs formed on the basis of perceptual appearances normally provide subjects with knowledge of the world (i.e., true warranted beliefs) because there is a reliable correlation between the appearances provided by perception (in appropriate conditions) and the first-order beliefs formed by the subject on the basis of those perceptual appearances. So, first-order beliefs formed on the basis of the appearances provided by perception (in appropriate conditions) will tend to be true. For instance, if I seem to perceive (in appropriate conditions) a sheep on the field and I form the first-order belief that there is a sheep on the field on that basis, my belief will normally be true. However, normally does not mean always. In every epistemic procedure of belief-formation there must be room for epistemic failure insofar as there is room for epistemic success. So, imagine that there is a bush on the field but (in appropriate perceptual conditions) I form the *false* first-order belief that there is a sheep on the field because the bush appears like a sheep to me (from my visual perspective and perceptual position). This is an example of epistemic failure in the epistemic process of belief-formation responsible for perceptual beliefs. Nothing more needs to be explained here to have an example of an epistemic failure in such a process.

However, things are different in the case of Moore's paradox. To explain Moore's paradox, it is not enough that an epistemic failure occurs in the first-person process of belief-formation (i.e., bypass procedure), it is also necessary that the subject *irrationally assents* to a Moore's sentence (i.e., "p, but I don't believe that p" or "p, but I believe that not-p") from the first-person deliberative perspective as a result of such an epistemic failure. However, it is not possible to describe a case in which a subject irrationally assents to a Moore's sentence from the first-person perspective because of an epistemic failure in the bypass procedure. The most similar thing to an example of a subject who has made the epistemic error in the bypass procedure that allegedly causes Moore's paradox is Tom's case. However, Tom cannot irrationally assent to "p, but I don't believe that p" or "p, but I believe that not-p" from the first-person deliberative perspective. So, even assuming (for the sake of argument) that Tom has made an epistemic error in the bypass procedure, he cannot be the example of a subject who gives rise to an instance of Moore's paradox because Moore's paradox requires that the subject irrationally assents to "p, but I don't believe that p" or "p, but I believe that not-p" from the first-person deliberative perspective and Tom cannot assent to this from the first-person deliberative perspective.

Indeed, from Fernández's account's perspective, the following account of Tom's case seems to follow. On the one hand, Tom has what he takes as good grounds to think that men and women are equal (for he doesn't find any good reason to think that men and women are not equal when deliberating from the first-person perspective), and so, he forms the first-order belief that men and women are equal. On the other hand, against what is described by the bypass procedure, he makes the epistemic error of forming the false second-order belief that he believes that men and women are not equal on the basis of the very same grounds. However, even if this could be an example of epistemic failure in the bypass procedure, this is not an example of Moore's paradox just yet. For we still need to describe a case in which Tom irrationally assents from a full first-person perspective to "Men and women are equal, but I believe that they are not" as a result of the epistemic failure that he has allegedly made in the application of the bypass procedure. And Tom cannot do that because he can only find out that he actually believes that men and women are not equal from the third-person perspective of self-inspection. On the one hand, Tom assents to "Men and women are equal" because he is supposed to have the true first-order belief that men and women are equal. On the other hand, Tom assents to "I believe that they are not" only because he finds from the third-person perspective of self-inspection that he acts in a way that is characteristic of the belief that men and women are not equal. As a result, when Tom can assent to "Men and women are equal, but I believe that are not", that assent is not an instance of Moore's paradox because no irrationality arises insofar as "I believe that they are not" can only be assented to from the third-person perspective.

Therefore, no example of Moore's paradox caused by an epistemic failure in the bypass procedure has been found so far because Tom's case is not an example of a subject who *irrationally* assents to a Moore's sentence from the *first-person perspective* as a result of an epistemic failure in the bypass procedure. Then, it seems that Moore's paradox cannot be explained by an epistemic failure in the bypass procedure of belief-formation. If Moore's paradox were the result of an epistemic failure in the bypass procedure, there would be cases of epistemic failure in the bypass procedure that give rise to an instance of a Moore's paradox (i.e., to the subject's irrational assent to a Moore's sentence from the first-person perspective). However, it is not possible to find a case of epistemic failure in the bypass procedure that gives rise to an instance of Moore's paradox (e.g., Tom cannot assent to a Moore's sentence from the first-person deliberative perspective even after having allegedly made an epistemic failure in the bypass procedure). As a result, Moore's paradox cannot be explained by claiming that it is the result of an epistemic failure in the bypass procedure.

Moran gives an account of Moore's paradox that is very similar to Fernández's account, adjusting the explanans to his agential account of Transparency and of first-person epistemic self-knowledge. According to Moran, Moore's paradox is the result of a failure in the first-person deliberative process responsible for Transparency and first-person self-knowledge. To assent to a Moore's sentence from the first-person perspective is irrational because it involves a failure in the first-person deliberation about whether p so that an *alienated belief* is caused. Alienated beliefs are beliefs that are not under the influence of the subject's reason-based deliberations, and so, they lack the property of self-awareness or self-reflection that is supposed to be responsible for first-person self-knowledge. Then, insofar as Moore's paradox involves the occurrence of an alienated belief, it also involves a lack of first-person self-knowledge.

What is interesting about Moran's account is that he seems to think that Tom-like cases[59] are actual examples of Moore's paradox. For, according to Moran, Tom's case is an example of a subject who has the alienated belief that men and women are not equal. Indeed, when Tom deliberates about whether men and women are equal, he concludes that they are. Since this belief is formed in a deliberative way, Tom is self-aware of the fact that he believes that men and women are equal. But when Tom self-inspects himself (or, in Moran's terminology, when he adopts a theoretical perspective about his own mental states), he finds out that he also has the alienated belief that men and women are not equal (a belief that is not self-reflective because it is not under the influence of his deliberations, and so, Tom is not self-aware of having that belief). As a result, Moran seems to think that Tom would exemplify a case of Moore's paradox if he assented to "Men and women are equal, but I believe that they are not" insofar as there is a conflict between what he judges from the deliberative first-person perspective and what he judges from the self-inspective third-person perspective. The failure of Transparency behind Moore's paradox is understood here as the conflict between the judgement made by the first-person deliberative perspective (i.e., "Men and women are equal") and the judgement made by the third-person perspective (i.e., "I believe that they are not"). This failure of Transparency is supposed to originate Moore's paradox because it is supposed to originate Tom's alienated belief that men and women are not equal.

However, this explanation of Moore's paradox doesn't work because it doesn't respect the basic features of the phenomenon. It faces the following problem: from the perspective that Moore's paradox arises (i.e., the deliberative perspective), Transparency doesn't fail; and from the perspective that Transparency fails (i.e., the self-inspective or theoretical perspective), Moore's paradox doesn't arise. So, Moore's paradox cannot be the result of a failure of Transparency, not even when Transparency is understood as agential deliberation based on reasons, as Moran does. Let's see the argument in detail. On the one hand, the first-person deliberative perspective is the perspective in which Moore's paradox is supposed to arise because it is the perspective from which it is irrational to assent to a Moore's sentence. However, from the first-person perspective, no failure of Transparency occurs in Moran's account. As Moran's himself says, subjects are not aware of having an alienated belief from the first-person perspective (i.e., subjects lack first-person self-knowledge of alienated beliefs), and so, when deliberating from the first-person deliberative perspective, Transparency doesn't

[59] Moran (1997, pp. 148-151; 2001, pp.77-83) discusses the example of the akratic gambler, which is similar to Tom's example, but it is built around intention instead of around belief.

fail even in subjects with alienated beliefs. Indeed, subjects with the alienated belief that p will answer the question "Do you believe that p?" in the same way as the question "Is p the case?" as long as they answer both questions from the first-person perspective. For instance, Tom is not aware of having the alienated belief that men and women are not equal from the first-person deliberative perspective. And so, from the first-person perspective, he will answer the questions "Do you believe that men and women are equal?" and "Are men and women equal?" in the same way: "Yes, men and women are equal" or "Yes, I believe that men and women are equal". Thus, no failure of Transparency occurs from the first-person perspective in which the phenomenon of Moore's paradox arises. On the other hand, the third-person theoretical or self-inspective perspective is the perspective from which Transparency fails in Moran's account. For it is by self-inspecting themselves from the third-person perspective that subjects can find out whether they have a certain alienated belief, assenting so to "p, but I don't believe that p" or to "p, but I believe that not-p". However, as we already know, Moore's paradox doesn't arise when subjects assent to a Moore's sentence (or to part of a Moore's sentence) from the third-person theoretical or self-inspective perspective because it is not irrational to do so. Thus, Moran's account doesn't actually explain Moore's paradox by claiming that it is the result of a failure of Transparency understood in an agential way: when Moore's paradox is supposed to arise (i.e., first-person perspective), Transparency doesn't fail (i.e., subjects answer both questions in the same way); and when Transparency fails (i.e., third-person perspective), Moore's paradox doesn't occur (i.e., assenting to a Moore's sentence is not irrational).

Therefore, epistemic accounts of Moore's paradox fail to explain the first desideratum (i.e., why it is sometimes irrational to assent to a Moore's sentence). Fernández proposes an account of Moore's paradox that implies that assenting from the first-person perspective to "p, but I don't believe that p" or to "p, but I believe that not-p" has to be a conceptually possible mistake of the subject, when it is not (as we know because no example can be described). And Moran proposes an account in which Moore's paradox is supposed to be explained by a failure of Transparency, but in which Moore's paradox and the failure of Transparency cannot occur at the same time (leaving so Moore's paradox unexplained). Notice that, insofar as no example of epistemic failure in the bypass procedure has been provided, the objection to Fernández's account is the same as the objection to Moran's account: when the failure of Transparency appears (i.e., third-person perspective), Moore's paradox doesn't arise; and when Moore's paradox arises (i.e., first-person perspective), no failure in the Transparency procedure appears (for the subject answers the questions “Do you believe that p?” and “Is p the case?” in the same way).

4.5 Semantic accounts of Moore’s paradox

Semantic accounts of Moore’s paradox claim that assenting to a Moore’s sentence from the first-person deliberative perspective is irrational because, in spite of appearances, it somehow hides a semantic contradiction. Linville & Ring (1991) claim that Moore’s paradox is explained because to assent to “I believe that p” from the first-person perspective is similar to assenting to “p”. Then, assenting to “p, but I believe that not-p” is similar to assenting to the straight contradiction “p, but not-p”, which explains why it is irrational to do so. However, this account of the paradox, as it stands, doesn’t account for the fact that Moore’s sentences can be true$_{rs}$ (i.e., the third desideratum of the paradox). For Linville & Ring don’t say anything about the possibility of a third-person self-inspective context in which it is possible to assent to a Moore’s sentence without that assent being irrational to make, and so, without giving rise to an instance of Moore’s paradox (e.g., Tom-like cases).

Fortunately, Heal (1994) develops a semantic account of Moore’s paradox that seems to explain why Moore’s sentences can have possible truth$_{rs}$-conditions. According to Heal, to *sincerely* think “I believe that p”, which is a self-ascription of belief, has a *performative* character. In the same way that saying “I promise to bring your book tomorrow” constitutes my promise to bring the book (ruling out infelicities incompatible with the realization of such a performative act, like being forced to say so), the sincere second-order thought “I believe that p” constitutes my belief that p (ruling out infelicities incompatible with the realization of such a performative act, like being under the effect of a drug). The qualification of “sincerity” is important in the case of believing. For saying “I promise that…” constitutes a promise regardless of whether the subject is sincere or not (i.e., regardless of whether the subject has the intention to fulfil the promise or not), but the second-order thought “I believe that p” is supposed to constitute the subject’s belief that p only when the subject is sincere in thinking that she believes that p. Thus, according to Heal, the sincere utterance “I believe that p” plays a double role when it is made from the first-person perspective: it is a self-ascription of belief (for it represents the subject as fulfilling the conditions for believing that p) and it is an assertion of the fact that p (for, due to its constitutive character, it expresses the first-order belief that p).

Then, Heal explains Moore's paradox by claiming that assenting to "p, but I believe that not-p" from the first-person perspective is irrational because it hides a semantic contradiction. When we see "I believe that not-p" as a self-ascription of belief, we see the sense in which Moore's sentences can be true because we see "I believe that not-p" as having the truth$_{rs}$-conditions of "Jesús believes that not-p" rather than the truth$_{rs}$-conditions of "p is not the case" (allegedly explaining in this way the third desideratum of Moore's paradox: why Moore's sentences can be true$_{rs}$). Here we are puzzled about why it is irrational to assent to "p, but I believe that not-p" from the first-person perspective. But when we see "I believe that not-p" as expressing the first-order belief that not-p, we see it as an alternative way to assent to "not-p", and so, as being contradictory with the first part of the sentence (i.e., "p"). Here we see why it is irrational to assent to "p, but I believe that not-p" from the first-person perspective: such assent hides, in one of the roles played by "I believe that not-p", the semantic contradiction "p, but not-p".

The advantage of Heal's account over Linville's & Ring's account is that it could explain why Moore's sentences can be true$_{rs}$ (third desideratum) without leaving unexplained why they are irrational to assent sometimes (first desideratum). Insofar as "I believe that p" plays the dual role of being a self-ascription of belief and an assertion of the fact that p, the sentence "p, but I believe that not-p" is both a contradiction and a sentence with possible truth$_{rs}$-conditions. However, both Heal's and Linville's & Ring's accounts inevitably fail to explain Moore's paradox in the case of "p, but I don't believe that p". Insofar as there is a conceptual difference between a self-ascription of *belief* and a self-ascription of *lack of belief*, "I don't believe that p" cannot be understood as being (always) tantamount to an assertion of not-p, and so, it cannot be understood as constituting (always) the first-order belief that not-p. As a result, no contradiction can be generated in the case of "p, but I don't believe that p" and Heal's and Linville's & Ring's semantic accounts cannot explain why it is sometimes irrational to assent to the Moore's sentence "p, but I don't believe that p", failing so to explain the second desideratum of Moore's paradox (i.e., that the two versions of the paradox should be appropriately explained).

Therefore, since no available account of Moore's paradox in the literature explains the four desiderata of the phenomenon of Moore's paradox in an appropriate way, it is in order to seek out a new account of Moore's paradox.

4.6 The behavioural-expressivist account of Moore's paradox

In this section, the semantic account of Moore's paradox that follows from the behavioural-expressivist account of Transparency is going to be developed. This semantic account of Moore's paradox claims that Moore's sentences are irrational to assent from the first-person deliberative perspective because Moore's sentences are self-contradictory (i.e., they don't have possible truth$_{rs}$-conditions) and self-contradictory-like (i.e., they don't have possible truth$_{nrs}$-conditions) when they are assented to from the first-person deliberative perspective. Afterwards, in the following section, it is going to be argued that this behavioural-expressivist account of Moore's paradox explains all the desiderata of a good account of Moore's paradox because it manages to avoid the problems faced by other semantic accounts currently available in the literature (i.e., being unable to explain why Moore's sentences can have truth$_{rs}$-conditions and/or being unable to explain both versions of Moore's paradox).

Epistemic accounts of Transparency consider that the questions "Do you believe that p?" and "Is p the case?" ask about different subject matters (i.e., whether the subject believes that p and whether p is the case, respectively) and that the former is transparent to the latter *when and only when* the subject answers to "Do you believe that p?" from the first-person deliberative perspective, acquiring so first-person epistemic self-knowledge. Then, according to epistemic accounts of Transparency, when the question "Do you believe that p?" is answered from the first-person deliberative perspective, the answer is a first-person avowal (e.g., "I believe that p") consisting in a *self-ascription of attitude* (e.g., belief, disbelief or lack of belief) made as the conclusion of a first-person deliberation about whether p. Since all accounts of Moore's paradox (with the exception of Linville's & Ring's account) consider that the irrationality of Moore's paradox arises because Moore's sentences are irrational to assent from the first-person deliberative perspective in spite of the fact that "p" is about whether p and that "I don't believe that p" or "I believe that not-p" is a *self-ascription of lack of belief* or a *self-*

ascription of the belief that not-p, all the accounts of Moore's paradox (with the exception of Linville's & Ring's account) assume an epistemic notion of Transparency.

By contrast, the behavioural-expressivist account of Transparency considers that the question "Do you believe that p?" can be meant both in a deliberative and in a self-ascriptive way. When it is meant in a deliberative way, the question "Do you believe that p?" asks about the fact that p (so that it is transparent to the question "Is p the case?") and the subject is supposed to answer with a *judgement about whether p* (which can take the linguistic form of a first-person avowal —e.g., "I believe that p"— or of an assertion —e.g., "p is the case"—) issued at the conclusion of a first-person deliberation about whether p. And when it is meant in a self-ascriptive way, the question "Do you believe that p?" asks about the subject's beliefs (so that it is not transparent to the question "Is p the case?") and it is supposed to be answered with a *self-ascription of attitude* (e.g., belief, disbelief or lack of belief) made as the conclusion of a third-person process of self-inspection based on evidence about one's own mental states. Therefore, from the behavioural-expressivist account of Transparency, it follows that Moore's paradox arises because of a conceptual mistake regarding the deliberative and the self-ascriptive uses of "I believe that p" (particularly, of "I don't believe that p" and "I believe that not-p"). When a Moore's sentence is assented from a full first-person deliberative perspective, it is irrational to assent because it is self-contradictory (i.e., it doesn't have possible truth$_{\mathrm{rs}}$-conditions) and self-contradictory-like (i.e., it doesn't have possible truth$_{\mathrm{nrs}}$-conditions) insofar as the two parts of the Moore's sentence (i.e., "p" and "I believe that not-p" or "I don't believe that p") are supposed to answer the deliberative question "Do you believe that p?". And when a Moore's sentence is partially assented to from the third-person self-inspective perspective (e.g., Tom-like cases), it is not irrational to assent to because it is not self-contradictory (i.e., it has possible truth$_{\mathrm{rs}}$-conditions) or self-contradictory-like (i.e., it has possible truth$_{\mathrm{nrs}}$-conditions): one part of the sentence (i.e., "p") answers the deliberative question "Do you believe that p?" and the other part of the sentence (i.e., "I believe that not-p" or "I don't believe that p") answers the self-inspective question "Do you believe that p?".

Since the account of the self-contradiction and of the self-contradiction-like characteristic of Moore's paradox that is going to be offered here is based on the relational and the non-relational concepts of truth that were explained in the last chapter, it might be useful to sketch this distinction again briefly. An expressive episode of mental state can be seen from two different perspectives: either as being a *presentation* of a mental state of the subject or as being *related* to an aspect of the world in an appropriate or inappropriate way (i.e., as having intentionality). For instance, the utterance “It is raining” can be seen either as presenting an aspect of my belief that it is raining or as being an assertion of the fact that it is raining. On the one hand, when an expressive episode is seen as being a presentation of a particular mental state of the subject, it can be either true$_{nrs}$ or false$_{nrs}$. An expressive episode is true$_{nrs}$ when it is presented as being *also in appearance* an expressive episode of the mental state that it is actually an episode of. For example, my utterance “It is raining” is true$_{nrs}$ if, as it appears to be, it is actually an episode of belief (and not of pretending to believe, for instance). By contrast, an expressive episode is false$_{nrs}$ when it is presented as being *in appearance* an expressive episode of a mental state different from the mental state that it is actually an episode of. For example, the utterance “It is raining” is false$_{nrs}$ if it is presented as an apparent episode of belief when it is actually an episode of pretending to believe. On the other hand, when an episode of expression is seen as being related to an aspect of the world in an appropriate or inappropriate way (i.e., as having intentionality), it can be either true$_{rs}$ or false$_{rs}$. An expressive episode is true$_{rs}$ when the appropriate relation of fit between its expressive content and the corresponding aspect of the world takes place. For instance, the utterance “It is raining” is true$_{rs}$ if it is a fact that it is raining. By contrast, an expressive episode is false$_{rs}$ when the appropriate relation of fit between its expressive content and the corresponding aspect of the world doesn’t take place. For instance, the utterance “It is raining” is false$_{rs}$ if the rain doesn’t take place. The expressive content of an episode of expression has a propositional kind of intentionality when it is actually or possibly related with a fact or state of affairs (e.g., that it is raining) rather than with an object (e.g., a friend). So, a *propositional content* is just a kind of expressive content; i.e., the kind of expressive content that is actually or possibly related with a certain fact or state of affairs (e.g., that it is raining) rather than with a certain object.

Once the non-relational and relational concepts of truth have been recalled, let’s see in order how behavioural expressivism can explain the two versions of the paradox, starting with the Moore’s sentence “p, but I believe that not-p” and moving on to “p, but I don’t believe that p” later on. On the one hand, assenting to “p, but I believe that not-p” from the first-person deliberative perspective is irrational because 1) it is self-contradictory-like in regard to its truth$_{nrs}$-conditions, and so, 2) it is self-contradictory in regard to its truth$_{rs}$-conditions as well. Assenting to “p, but I believe that not-p” from the first-person deliberative perspective has 1) impossible truth$_{nrs}$-conditions (i.e., it is self-contradictory-like) because “p, but I believe that not-p” is supposed to be a single expressive episode (i.e., a single act of assent) that is presented as having *in appearance* the expressive content of two incompatible attitudes at once: the belief that p and the belief that not-p. Indeed, the belief that p and the belief that not-p are incompatible attitudes (i.e., attitudes that cannot be held by a subject at the same time) insofar as a subject cannot answer the first-person deliberative question "Do you believe that p?" with the judgement that p and with the judgement that not-p at the same time. Thus, assenting to "p, but I believe that not-p" from the first-person deliberative is an act of assent that presents itself as having the *apparent* expressive content of the following incompatible attitudes: the belief that p (for "p" is supposed to be the judgement that p, and so, "p" is presented as being in appearance an episode of the belief that p) and the belief that not-p (for "I believe that not-p" is supposed to be the judgement that not-p, and so, "I believe that not-p" is presented as being in appearance an episode of the belief that not-p). As a result, assenting to "p, but I believe that not-p" doesn't have possible truth$_{nrs}$-conditions (i.e., it is self-contradictory-like) because it is an act of assent that appears to express the incompatible attitudes *belief that p* and *belief that not-p*, and so, it is an act of assent that doesn't actually have any expressive content because it doesn't express any mental state at all.

Moreover, the act of assenting to "p, but I believe that not-p" from the first-person deliberative perspective is 2) self-contradictory as well (i.e., it doesn't have possible truth$_{rs}$-conditions) because it is a *nonsense* that doesn't say anything about the world (neither true$_{rs}$ nor false$_{rs}$). The fact that assenting to "p, but I believe that not-p" is self-contradictory (i.e., it doesn't have possible truth$_{rs}$-conditions) is dependent of the fact that it is self-contradictory-like (i.e., it doesn't have possible truth$_{nrs}$-conditions). For, since the assent to "p, but I believe that not-p" doesn't have any *expressive content* because it doesn't express any mental state (insofar as it doesn't have possible truth$_{nrs}$-conditions), "p, but I believe that not-p" doesn't have any intentionality or propositional content either, and so, it doesn't say anything about the world (neither true$_{rs}$ nor false$_{rs}$). Then, since assenting to "p, but I believe that not-p" is a nonsense that doesn't say anything about the world, it is self-contradictory because it doesn't have possible truth$_{rs}$-conditions.

On the other hand, when one of the parts of the Moore's sentence that is assented to answers the self-ascriptive question "Do you believe that p?" from the third-person perspective of self-inspection, assenting to "p, but I believe that not-p" is not an irrational act because there isn't any self-contradiction or self-contradiction-like involved (i.e., it has both possible truth$_{rs}$-conditions and possible truth$_{nrs}$-conditions). With "p" the subject answers the deliberative question "Do you believe that p?" making the judgement that p is the case as the conclusion of a first-person deliberation about whether p. And with "I believe that not-p" the subject answers the self-ascriptive question “Do you believe that p?” making a self-ascription of the belief that not-p as the conclusion of a process of self-inspection about whether she *believes* that p. Since having the first-order belief that p is not incompatible with having the second-order belief that one believes that not-p (for the first-order belief has to do with the question “Do you believe that p?” meant in a deliberative way and the second-order belief has to do with the question “Do you believe that p?” meant in a self-ascriptive way), no self-contradiction or self-contradiction-like arises. In this conversational context, “p” is both in appearance and in reality an expressive episode of the first-order belief that p that consists in the judgement that p is the case, and “I believe that not-p” is both in appearance and in reality an expressive episode of the *second-order* belief that I believe that not-p that consists in a self-ascription of the belief that not-p (i.e., in the judgement that I have the first-order belief that not-p). Hence, in this conversational context, “p, but I believe that not-p” have both possible truth$_{nrs}$-conditions and possible truth$_{rs}$-conditions.

Tom is an example of a subject who can assent to the Moore’s sentence “Men and women are equal, but I believe that they are not” (i.e., “p, but I believe that not-p”) without that assent being irrational to make. Since Tom assents to “Men and women are equal” to answer the deliberative question “Do you believe that men and women are equal?” from the first-person deliberative perspective (i.e., deliberating on the basis of evidence about gender equality), “Men and women are equal” is both in appearance and in reality an expressive episode of the belief that men and women are equal that consists in judging that they are. And since Tom assents to “I believe that men and women are not equal” from the third-person perspective of self-inspection (on the basis of the evidence about his mental states that supports the fact that he acts as if he believed that men and women are not equal), “I believe that men and women are not equal” is both in appearance and in reality an expressive episode of the second-order belief that he believes that men and women are not equal that consists in a self-ascription of that sexist belief (i.e., in the judgement that he has that sexist belief). Indeed, Tom finds out by self-inspection that he acts like a person who believed that men and women are not equal would act (in spite of judging that men and women are equal when deliberating about gender equality from the first-person perspective), and so, he judges by self-inspection that he has the first-order belief that men and women are not equal (with the self-ascription of belief: “I believe that they are *not* equal”)[60]. Therefore, Tom’s assent to “Men and women are equal,

[60] In fact, as we know from the last chapter, Tom doesn’t actually have the *belief* that men and women are not equal but the *self-deceived belief* that men and women are equal. However, Tom doesn’t need to know that,

but I believe that they are not" is not irrational to make because there isn't any self-contradiction or self-contradiction-like involved. In that context, assenting to "p, but I believe that not-p" has both possible truth$_{nrs}$-conditions and possible truth$_{rs}$-conditions.

Moving on to the other version of the paradox, the Moore's sentence "p, but I don't believe that p" is explained by the behavioural-expressivist account in the following way. On the one hand, assenting to "p, but I don't believe that p" from the first-person deliberative perspective is irrational because 1) it is self-contradictory-like in regard to its truth$_{nrs}$-conditions, and so, 2) it is self-contradictory in regard to its truth$_{rs}$-conditions as well. Assenting to "p, but I don't believe that p" from the first-person deliberative perspective has 1) impossible truth$_{nrs}$-conditions (i.e., it is self-contradictory-like) because "p, but I don't believe that p" is supposed to be a single expressive episode (i.e., a single act of assent) that is presented as having *in appearance* the expressive content of two incompatible attitudes at once: the belief that p and the lack of belief about p. Indeed, the belief that p and the lack of belief about p are incompatible attitudes (i.e., attitudes that cannot be held by a subject at the same time) insofar as a subject cannot answer the first-person deliberative question "Do you believe that p?" with the judgement that p and with a suspension of judgement about p at the same time. Thus, assenting to "p, but I don't believe that p" from the first-person deliberative perspective is an act of assent that presents itself as having the *apparent* expressive content of the following incompatible attitudes: the belief that p (for "p" is supposed to be the judgement that p, and so, "p" is presented as being in appearance an episode of the belief that p) and lack of belief about p (for "I don't believe that p" is supposed to be a suspension of judgement about p, and so, "I don't believe that p" is presented as being in appearance an episode of lack of belief about p). As a result, assenting to "p, but I don't believe that p" doesn't have possible truth$_{nrs}$-conditions (i.e., it is self-contradictory-like) because it is an act of assent that appears to express the incompatible attitudes *belief that p* and *lack of belief about p*, and so, it is an act of assent that doesn't actually have any expressive content because it doesn't express any mental state at all.

Moreover, the act of assenting to "p, but I don't believe that p" from the first-person deliberative perspective is 2) self-contradictory as well (i.e., it doesn't have possible truth$_{rs}$-conditions) because it is a *nonsense* that doesn't say anything about the world (neither true$_{rs}$ nor false$_{rs}$). The fact that assenting to "p, but I don't believe that p" is self-contradictory (i.e., it doesn't have possible truth$_{rs}$-conditions) is dependent of the fact that it is self-contradictory-

especially taken into account that self-deceived beliefs can be easily confused with regular beliefs from the third-person process of self-inspection.

like (i.e., it doesn't have possible truth$_{nrs}$-conditions). For, since the assent to "p, but I don't believe that p" doesn't have any expressive content because it doesn't express any mental state (insofar as it doesn't have possible truth$_{nrs}$-conditions), the assent to "p, but I don't believe that p" doesn't have any intentionality or propositional content either, and so, it doesn't say anything about the world (neither true$_{rs}$ nor false$_{rs}$). Then, since assenting to "p, but I don't believe that p" is a nonsense that doesn't say anything about the world, it is self-contradictory because it doesn't have possible truth$_{rs}$-conditions.

On the other hand, when one of the parts of the Moore's sentence is assented to in order to answer the self-ascriptive question "Do you believe that p?" from the third-person perspective of self-inspection, assenting to "p, but I don't believe that p" is not an irrational act for the same reason that we saw in the other version of the paradox: there isn't any self-contradiction or self-contradiction-like involved, and so, it has both possible truth$_{rs}$-conditions and possible truth$_{nrs}$-conditions. With "p" the subject answers the deliberative question "Do you believe that p?" making the judgement that p is the case as the conclusion of a first-person deliberation about whether p. And with "I don't believe that p" the subject answers the self-ascriptive question "Do you believe that p?" making a self-ascription of lack of belief about p as the conclusion of a process of self-inspection about whether she *believes* that p. Since having the first-order belief that p is not incompatible with having the second-order belief that one doesn't have any belief about p (for the first-order belief has to do with the question "Do you believe that p?" meant in a deliberative way and the second-order belief has to do with the question "Do you believe that p?" meant in a self-ascriptive way), no self-contradiction or self-contradiction-like arises. In this conversational context, "p" is both in appearance and in reality an expressive episode of the first-order belief that p that consists in the judgement that p is the case, and "I don't believe that p" is both in appearance and in reality an expressive episode of the *second-order* belief that I don't have any belief about p that consists in a self-ascription of lack of belief about p (i.e., in the judgement that I don't have any belief about p). Hence, in this conversational context, "p, but I believe that not-p" have both possible truth$_{nrs}$-conditions and possible truth$_{rs}$-conditions.

Again, Tom's case could be constructed as an example of a subject who assents to the Moore's sentence "Men and women are equal, but I don't believe that they are" (i.e., "p, but I don't believe that p") without that assent being irrational to make. On the one hand, Tom assents to "Men and women are equal" to answer the deliberative question "Do you believe that men and women are equal?" from the first-person deliberative perspective (i.e., deliberating on the basis of evidence about gender equality), and so, "Men and women are equal" is both in appearance and in reality an expressive episode of the belief that men and women are equal that consists in the judgement that they are equal. On the other hand, imagine that this time, when Tom self-inspects himself about whether he believes that men and women are equal, he finds some behavioural evidence supporting the fact that he believes that men and women are equal (e.g., he has publicly defended the idea that men and women are equal multiple times in the past) and some behavioural evidence supporting the fact that he believes that men and women are not equal (i.e., he tends to rely more on men for intellectual tasks, he get nervous when he gets in a car driven by a woman, and so on). So, Tom ends the process of self-inspection this time judging that he doesn't believe neither that men and women are equal nor that they are not equal, for he acts in a way that is incompatible with having either of those two beliefs. A person who believed that men and women are equal wouldn't act in a sexist way, and a person who believed that men and women are *not* equal wouldn't publicly defend the opposite. As a result, when answering the self-ascriptive question "Do you believe that men and women are equal?" from the third-person perspective of self-inspection, Tom answers with the self-ascription of lack of belief "I don't believe that men and women are equal", which is both in appearance and in reality an expressive episode of the second-order belief that he doesn't have any belief about gender equality. Therefore, Tom's assent to "Men and women are equal, but I don't believe that they are" is not irrational to make because there isn't any self-contradiction or self-contradiction-like involved. In that context, "p, but I don't believe that p" has both possible truth$_{nrs}$-conditions and possible truth$_{rs}$-conditions.

Thus, from the behavioural-expressivist account of Transparency, together with the behavioural-expressivist notions of relational and non-relational truth, follows a semantic account that explains Moore's paradox claiming that when Moore's sentences are irrational to assent, they are self-contradictions and self-contradictions-like, and when Moore's sentences are not irrational to assent, they are not self-contradictions nor self-contradictions-like. In the next section, it is going to be argued that the behavioural-expressivist account of Moore's paradox, unlike the rest of accounts currently available in the literature, is able to explain all the desiderata that a good account of Moore's paradox should be able to explain.

4.7 Accounting for the desiderata of Moore's paradox

The behavioural-expressivist account of Moore's paradox proposed here explains the four desiderata of a good account of Moore's paradox: in the following way:

1) *Moore's sentences are sometimes irrational to assent to*. The behavioural-expressivist account of Moore's paradox explains that Moore's sentences are sometimes irrational to assent to because it claims that Moore's sentences are self-contradictions-like and self-contradictions when and only when they are assented to from the first-person deliberative perspective; i.e., when and only when they are supposed to answer the question "Do you believe that p?" meant in a deliberative way.

2) *Two versions of the paradox: "p, but I don't believe that p" and "p, but I believe that not-p"*. The behavioural-expressivist account of Moore's paradox explains why assenting to a Moore's sentence from the first-person deliberative perspective is irrational both in the case of "p, but I don't believe that p" and "p, but I believe that not-p" while respecting the conceptual differences between them (particularly, between "I don't believe that p" and "I believe that not-p"). Assenting to "p, but I don't believe that p" from the first-person deliberative perspective would be like expressing the two incompatible attitudes *belief that p* and *lack of belief that p* in the same episode of expression. By contrast, assenting to "p, but I believe that not-p" from the first-person deliberative perspective would be like expressing the two incompatible attitudes *belief that p* and *belief that not-p*. Therefore, since the assents to "p, but I don't believe that p" and "p, but I believe that not-p" don't have any expressive content because, in spite of appearances, they don't express any mental state (i.e., they are self-contradictory-like), they don't have any intentionality or propositional content either, and so, they cannot say something true$_{rs}$ about the world (i.e., they are self-contradictory).

   This is one of the desiderata that Heal's and Linville's & Ring's semantic accounts of Moore's paradox failed to explain. As we saw, their semantic accounts could explain why assenting to "p, but I believe that not-p" is irrational because they claim that assenting to "I believe that not-p" involves assenting to "not-p", and so, that assenting to "p, but I believe that not-p" is self-contradictory because that act of assent doesn't have possible truth$_{rs}$-conditions (i.e., it is like assenting to "p" and "not-p"). However, they cannot explain why assenting to "p, but I don't believe that p" is irrational because assenting to "I don't believe that p" is not always equivalent to assenting to "not-p" (i.e., it can be a suspension of judgement or a self-ascription of lack of belief). From the perspective of the behavioural-expressivist account, it is not surprising that traditional semantic accounts of Moore's paradox cannot explain why assenting to "p, but I don't believe that p" can be just as irrational as assenting to "p, but I believe that not-p". Self-contradictions in the relational sense (i.e., lack of possible truth$_{rs}$-conditions) are logically dependent of self-contradictions in the non-relational sense (i.e., lack of possible truth$_{nrs}$-conditions) because the *intentionality* of an episode of expression (on which the truth$_{rs}$-value of an episode of expression depends) is logically dependent of the expressive content of that episode of expression (on which the truth$_{nrs}$-value of an episode of expression depends). Since semantic accounts of Moore's paradox other than behavioural expressivism don't have available the notion of non-relational truth, they understand the concept of contradiction in the wrong way: as assertions that say something about the world (so that they have intentionality), but that don't have possible truth$_{rs}$-conditions because what they say about the world is an impossible fact (e.g., that it is raining and that it isn't raining). As a result, they cannot explain what there can be of self-contradictory in the Moore's sentence "p, but I don't believe that p": it cannot be argued that such an assertion says something impossible about the world insofar as "I don't believe that p" doesn't say anything about the fact that p (i.e., it is supposed to be either a self-ascription of lack of belief or an expression of suspension of judgement about p).

By contrast, the behavioural-expressivist account appropriately explains why assenting to "p, but I don't believe that p" can be irrational because it understands the concept of contradiction in the right way: the truth$_{rs}$-value of an episode of expression is logically dependent of its expressive content because the expressive content of an episode of expression is the condition of its intentionality or propositional content. So, assenting to "p, but I don't believe that p" from the first-person deliberative perspective is self-contradictory in the relational sense because it is self-contradictory in the non-relational sense. If it is self-contradictory in the non-relational sense (i.e., if it doesn't have possible truth$_{nrs}$-conditions), it doesn't have any expressive content. And if it doesn't have any expressive content, it doesn't have any intentionality or propositional content either. As a result, it is self-contradictory in the relational sense as well because it is a *nonsense*: it doesn't have possible truth$_{rs}$-conditions because it doesn't say anything about the world (neither possible nor impossible).

3) *Moore's sentences can have possible truth-conditions*. The behavioural-expressivist account of Moore's paradox explains that Moore's sentences can have possible truth-conditions (both in the relational and in the non-relational sense) because it claims that assenting to a Moore's sentence involves different commitments when the Moore's sentence is fully assented to as an answer to the question "Do you believe that p?" meant in a deliberative way, and when the Moore's sentence is partially assented to as an answer to the question "Do you believe that p?" meant in a self-ascriptive way. When a Moore's sentence is fully assented to as an answer to the question "Do you believe that p?" meant in the deliberative way, it is self-contradictory and self-contradictory-like. And when a Moore's sentence is partially assented to as an answer to the question "Do you believe that p?" meant in a self-ascriptive way, it has both possible truth$_{nrs}$-conditions and possible truth$_{rs}$-conditions. This is another of the desiderata that Linville & Ring (1991) didn't explain because their semantic account doesn't describe any sense in which the assent to a Moore's sentence has possible truth-conditions.

4) *Moore's paradox can arise both when asserting a Moore's sentence aloud and when judging the content of a Moore's sentence silently in thought*. It is clear that the behavioural-expressivist account of Moore's paradox shouldn't have any problem explaining why it is irrational to assert aloud a Moore's sentence from the first-person deliberative perspective. Since it is clear that assertions are *linguistic expressions*, it is clear how the act of asserting a Moore's sentence from the first-person perspective can be a self-contradictory-like and a self-contradictory *episode*

of expression. However, it remains to be seen how the behavioural-expressivist account of Moore's paradox can explain why the act of judging the content of a Moore's sentence silently in thought from the first-person deliberative perspective is irrational as well. So, how can the behavioural-expressivist account explain the irrationality that arises when judging the content of a Moore's sentence silently in thought from the first-person deliberative perspective? Are judgements in thought *expressive episodes* as well?

The behavioural-expressivist account of Moore's paradox proposed here considers that silent judgements in thought are *expressive episodes* of mental states with the same right as aloud assertions[61]. The reason why that is so is that *thinking* a content (e.g., [I have to go for groceries tomorrow][62]) is nothing over and above *suppressed saying*; namely, it is *like* saying something (e.g., "I have to go for groceries tomorrow") without moving your mouth and without issuing any sound. As a result, silently judging in thought [I have to go for groceries tomorrow] and asserting aloud "I have to go for groceries tomorrow" are the *same kind of expressive episodes* of my belief that I have to go for groceries tomorrow because they occupy the same kind of positions in the context of the temporal expressive pattern of my belief that I have to go for groceries tomorrow: the position of an act of assent or commitment to the content that I have to go for groceries tomorrow. So, the difference between my silent judgement [I have to go for groceries tomorrow] and my aloud assertion "I have to go for groceries tomorrow" is that they are different types of *instantiations* of the same kind of *expressive episode* of mental state. On the one hand, they are the same kind of expressive episodes of belief because they are an act of assent or commitment to the content [I have to go for groceries tomorrow] that takes place in the context of the expressive pattern of my belief that I have to go for groceries tomorrow. On the other hand, they are two different types of instantiations of the same kind of expressive episode because they are instantiated

[61] The idea that thoughts are expressive as well is not new, it was already endorsed by Finkelstein:

> "For my part, I do find it natural to use the word 'expression' to refer even to thoughts. Imagine a man who is trying to tiptoe into bed so as not to disturb his sleeping wife. While thus engaged, he stubs a toe on one of the bed's metal legs. He feels the impulse to cry out but suppresses it and only thinks, 'Ow, ow, ow, ow; that *really* hurts.' I'd call this thought an expression of pain." (Finkelstein, 2008, p.112).

[62] Brackets will be used henceforth to mark when a content is thought silently rather than uttered aloud.

in *different sets of expressive vehicles*: while a silent judgement is an expressive episode (i.e., an act of assent) that is instantiated in vehicles of expression such as a facial expression, a certain demeanour or a bodily posture, an aloud assertion is an expressive episode (i.e., an act of assent) that is *also* instantiated in the aloud utterance "I have to go for groceries tomorrow" (i.e., an additional vehicle of expression that is lacking in silent judgements).

That judging a particular content silently and asserting that particular content aloud are the same kind of expressive episodes (instantiated in different sets of vehicles of expression) is shown by the fact that they occupy the same kind of positions in the expressive pattern of a mental state. In a first stage of the socialization process, subjects learn to express their mental states linguistically, *saying things aloud* about an aspect of the world in the context of a community and a form of life. However, in a second stage of their socialization process, once they have to some degree mastered the ability to linguistically express their mental states saying things aloud about an aspect of the world, subjects perfect that ability to a greater degree and they become capable of *suppressed saying*; i.e., of producing expressive episodes of mental states that are *like* saying something about an aspect of the world but without moving their mouth or issuing any sound; namely, they become capable of expressing their mental states (e.g., in a facial expression, in a demeanour, in a bodily posture, etc.) silently thinking about an aspect of the world. As a result, both thinking and talking are manifestations of the same ability (i.e., of the ability to express one's own mental states assenting or endorsing commitments to aspects of the world), with the qualification that thinking or *suppressed saying* is the result of improving that ability to a greater degree than the degree that is necessary for talking aloud. Therefore, silently judging [p] is the *same kind* of expressive episode as asserting "p" aloud (i.e., an act of assent or commitment), although they are instantiated in different sets of vehicles of expression (i.e., only in facial expressions, demeanours or bodily postures vs. *also* in aloud utterances). In both cases, these vehicles of expression acquire the expressive content of an act of assent or commitment to the content [p] because of the kind of position that they occupy in the context of the expressive pattern of my belief that p (i.e., the position of an act of assent or commitment to the content [p]), and so, the expressive content of these vehicles of expression includes the content [p], regardless of whether the aloud utterance "p" is a vehicle of expression among the vehicles of expression in which the expressive episode (i.e., the act of assent or commitment to [p]) is instantiated on the given occasion or not.

In order to give plausibility to the idea that thinking and talking are manifestations of the same ability to express one's own mental states endorsing commitments about an aspect of the world (being so constitutive of the same kind of expressive episodes instantiated in different vehicles of expression), one can describe the analogous example of how we learn the ability to read. Learning the ability to read is a process that includes many stages. Firstly, the subject practices the ability to pronounce aloud what she thinks is written down on the sheet, so she can check with the teacher if she is reading the word or the sentence properly. Secondly, after practice, the subject is able to read sentences or paragraphs whispering to herself, reading aloud just when she wants to check whether she is doing it right. Thirdly, the subject is able to read only by moving her lips as if she were talking, but without issuing any audible sound. And finally, she is able to read without saying anything aloud, without whispering and without moving her lips. It is at this point that we might say that the subject has mastered the ability to read to the degree of a normal literate person. However, even if the subject is now able to read silently and without moving her lips, she is exercising the same ability that she was exercising at the beginning of the process, only perfected and mastered after a long period of practice.

Therefore, to silently think or judge the content of a Moore's sentence from the first-person deliberative perspective is irrational because it is constitutive of the same kind of expressive episode as to assert aloud a Moore's sentence from the first-person deliberative perspective, and so, it has the same impossible truth$_{nrs}$-conditions (i.e., it is self-contradictory-like) and the same impossible truthn$_{rs}$-conditions (i.e., it is self-contradictory). For judging a Moore's sentence silently in thought from the first-person deliberative perspective is nothing over and above assenting to the content of a Moore's sentence as if the subject were asserting aloud a Moore's sentence from the first-person perspective but without moving her mouth and without issuing any sound (*suppressed saying*), and so, we are dealing with the same kind of expressive episodes in both cases (although instantiated in different vehicles of expression: a facial expression, a demeanour or a bodily posture vs. *also* a sentence pronounced aloud).

Therefore, the semantic account of Moore's paradox proposed here, which follows from the behavioural-expressivist account of Transparency, seems to be the best explanation of Moore's paradox among the accounts currently available in the literature: it is the only account that explains the four desiderata of a good account of Moore's paradox.

4.8 An objection to the behavioural-expressivist account of Moore's paradox

An intuitive objection to the account of Moore's paradox proposed here is the following. If to silently think or judge the content of a Moore's sentence from the first-person perspective is supposed to be irrational because it is supposed to be a self-contradictory-like and a self-contradictory *expressive episode*, just like saying aloud "p, but I believe that not-p" or "p, but I don't believe that p", that self-contradictory-like and self-contradictory expressive content must be publicly available to other people's perception in the subject's facial expression, demeanour or bodily posture, which are supposed to be the kind of vehicles of expression that constitute the expressive episode of silently judging the content of a Moore's sentence. However, even if I can perceive from the sense of hearing the self-contradictory-like and self-contradictory content of your assertion "p, but I believe that not-p" or "p, but I don't believe that p" when it is pronounced aloud (vehicles of expression) in a first-person deliberative context, I can't perceive the alleged self-contradictory-like and self-contradictory content of your first-person deliberative judgement "p, but I believe that not-p" or "p, but I don't believe that p" in your facial expression, bodily posture or demeanour (vehicles of expression) when you make that judgment silently in thought. Hence, it seems that your silent judgement of a Moore's sentence cannot be an *episode of expression* because it is not an episode of mental state available to other people's perception, and so, that the behavioural-expressivist account of Moore's paradox doesn't actually explain why it is irrational to silently judge in thought a Moore's sentence from the first-person deliberative perspective. In a nutshell, the objection is the following: 1) episodes of expression are bits of publicly available expressive behaviour; 2) silent judgements are not bits of publicly available expressive behaviour because they cannot be perceived by others in one's own facial expressions, bodily postures or demeanours (vehicles of expression); 3) hence, silent judgements are not episodes of expression, and so, the behavioural-expressivist account of Moore's paradox cannot plausibly explain why the act of judging silently in thought the content of a Moore's sentence from the first-person deliberative perspective is irrational.

This objection is an aspect of a more general objection against the behavioural-expressivist conception of mental states. It is obvious that I cannot perceive people's silent judgements from just looking at their faces and bodies (i.e., perceiving their current expressive vehicles or current expressive behaviour), and so, the objection continues, it is obvious that mental states cannot be identical to *patterns of expressive episodes* because being an expressive episode requires being publicly available to other people's perception, and silent judgements are not publicly available to other people's perception. So, insofar as silent judgements are not publicly available to other people's perception, silent judgements (e.g., the act of silent assent to [The Earth goes around the Sun]) must be *private episodes* of mental states (e.g., of my belief that the Earth goes around the Sun) rather than *expressive episodes* of mental states. In a nutshell, the objection goes as follows: 1) expressive episodes of mental states must be publicly available to other people's perception in order to be truly expressive; 2) there are private episodes of mental states (e.g., silent judgements) not available to other people's perception; 3) hence, mental states cannot be identical to patterns of expressive episodes, and so, the behavioural-expressivist conception of mental states is mistaken.

The answer that I propose to this objection is that silent judgements (e.g., the act of assent to [The Earth goes around the Sun]) are part of the expressive content of people's facial expressions, bodily postures or demeanours at the time in which they silently make the judgement. However, we cannot perceive people's silent judgements in their facial expressions, bodily postures or demeanours because the *appropriate perceptual conditions* to perceive the *whole expressive content* of the subjects' vehicles of expression (facial expressions, demeanours, bodily postures, etc.) cannot take place when the aloud utterance of the (silent) judgement is lacking (because it is a vehicle of expression that has been suppressed from the expressive episode; i.e., from the act of assent to [The Earth goes around the Sun]). The reason why the appropriate perceptual conditions cannot take place is that facial expressions, demeanours or bodily postures are very *unclear* expressive vehicles of those expressive episodes of mental states that consist in silent judgements (i.e., in silent acts of assent to a content). By contrast, when an expressive episode of mental state consisting in a judgment (e.g., the assent to the content [The Earth goes around the Sun]) has an aloud utterance (e.g., "The Earth goes around the Sun") among the vehicles of expression that constitute the expressive episode in question, we can normally perceive the whole expressive episode or judgment (e.g., the assent to the content [The Earth goes around the Sun]) in the aloud utterance (vehicle of expression) because the *appropriate perceptual conditions* to perceive the whole expressive content of such a *clear* vehicle of expression of judgements normally take place. This type of expressive episodes (i.e., judgements or acts of assent that are *also* constituted by aloud utterances) are called *assertions*. Thus, it is not that there are *private* episodes of mental states. By contrast, all episodes of mental states (including both silent judgements and assertions) are *expressive* episodes. But the perceptual conditions necessary to perceive the expressive content of a silent judgement in someone's facial expression, bodily posture or demeanour cannot take place.

Indeed, (non-empty) vehicles of expression are both material items and expressive items at the same time, and so, they have different perceptual conditions depending on whether they are perceived as material items or as expressive items. On the one hand, when vehicles of expression are perceived as material items, the perceptual conditions involved are the same as the perceptual conditions involved in perceiving any other material item (e.g., a table); i.e., physical factors such as the lighting conditions, the distance between the object and me, the volume of the sound and the distance between the source and me, whether the perceiving subject has uncorrected myopia or not, etc. On the other hand, when vehicles of expression are perceived as expressive items with a certain expressive content, there are additional perceptual conditions which are exclusive of perceiving an *expressive content*. These perceptual conditions are things such as the *degree of clarity* with which the vehicle of expression presents its expressive content, the *context* in which the expressive episode takes place, the *portion* of the expressive pattern of mental state that the perceiving subject is able to perceive, how much the perceiving subject *knows the other person* and her particular ways of expression, etc. For instance, I can perceive anger in someone's facial expression of anger (i.e., expressive vehicle) if there is enough light to perceive the facial expression of the subject (i.e., perceptual condition of the facial expression as a material item), and also, if the facial expression is a *clear* facial expression of anger rather than a subtle or unclear facial expression of anger, if I've perceived enough of the *context* in which the subject's facial expression of anger takes place (e.g., that someone tried to steal his wallet), if I've perceived a large enough *portion* of the expressive pattern of the subject's anger (e.g., if I keep watching him for a few seconds and he starts to complain about the pickpocket while keeping the same facial expression), and so on. The latter conditions (i.e., clarity of presentation of the expressive content, context of the expressive episode, portion of the expressive pattern of mental state perceived, and knowledge of the subject's idiosyncrasy) are all perceptual conditions of the expressive content of the facial expression as an expressive item (i.e., of the facial expression as an expression of anger).

Thus, the reason why it is not possible to perceive other people's silent judgements only in their facial expressions, bodily postures or demeanours is that these are very *unclear vehicles of expression* of silent judgements or acts of assent, and so, they constitute expressive episodes (i.e., silent judgements or acts of assent) whose expressive content is presented in a very *unclear* way[63] to other people's perception. As a result, the *appropriate perceptual conditions* to perceive the whole expressive content of an expressive episode consisting in a silent judgement (e.g., the act of assent to [The Earth goes around the Sun]) only in facial expressions, bodily postures or demeanours cannot take place. By contrast, when an expressive episode consisting in a judgement (e.g., the act of assent to [The Earth goes around the Sun]) is constituted, not only by unclear vehicles of expression of judgements (like facial expressions, demeanours or bodily postures), but also by clear vehicles of expression of judgements (like aloud utterances), the expressive episode (e.g., the judgement or act of assent) can present its expressive content in a clear way and the appropriate perceptual conditions can take place: it is possible to perceive the expressive content of the episode of mental state in the expressive vehicles of the subject because the aloud utterance (i.e., a clear expressive vehicle of judgements or acts of assent) is among the expressive vehicles that constitute the expressive episode in question. Again, this kind of expressive episodes (i.e., judgements or acts of assent that are *also* constituted by aloud utterances) are called *assertions*.

In this reply to the objection against the behavioural-expressivist notion of mental states there are two aspects involved, and they need to be tackled in more detail: an *ontological aspect* (i.e., what silent thoughts and silent judgements are as expressive episodes, namely, as content of some vehicles of expression) and an *epistemological aspect* (i.e., how we can *know* the expressive content of someone else's vehicles of expression when that expressive content includes silent thoughts or silent judgements if we cannot perceive those silent thoughts or silent judgements). Firstly, let's clarify the ontological aspect with an example. Suppose that I am *silently* deliberating about whether the Earth goes around the Sun and I reach the conclusion

[63] Notice that an expressive content presented in an *unclear way* is not the same as an expressive content presented in a *misleading way*. The latter involves falsehood in the non-relational sense of truth, while the former doesn't.

[The Earth goes around the Sun]. The ongoing deliberation are silent thoughts and the conclusion reached at the end is a silent judgement. They are *silent* because they are episodes of “talking” without moving the mouth and without emitting any sound, and so, they are *expressive episodes* either of my ongoing deliberation or of my newly formed belief that the Earth goes around the Sun. However, if those silent thoughts and silent judgement are expressive episodes, *where are* they? They are an expressive content included in the content of my expressive behaviour over the period of time that my deliberation and its conclusion occur: an expressive content included in the expressive content of my facial expression, of my bodily posture or of my demeanour (i.e., of my vehicles of expression). The expressive content of those vehicles of expression (my facial expression, my bodily posture, my demeanour, etc.) *includes* the particular content of the thoughts that constitute my deliberation about whether the Earth goes around the Sun and the particular content of the silent judgement that constitutes my conclusion of that deliberation because those vehicles of expression (my facial expression, my bodily posture, my demeanour, etc.) occupy exactly the same *kind of places* in the context of the *expressive pattern* of my deliberation about whether the Earth goes around the Sun and in the context of the *expressive pattern* of my newly formed belief that the Earth goes around the Sun as the relevant pronounced utterances (vehicles of expression) could have occupied.

Therefore, my facial expression, my bodily posture or my demeanour over the course of the time that I am deliberating are expressive vehicles of my ongoing deliberation whose expressive content includes the ongoing thoughts that constitute my deliberation. Thus, my facial expression, my bodily posture or my demeanour are among the vehicles of expression that constitute the expressive episodes of my ongoing deliberation. Also, my facial expression, my bodily posture or my demeanour when I reach the conclusion of that deliberation are expressive vehicles whose expressive content includes my silent judgement [The Earth goes around the Sun]. Thus, my facial expression, my bodily posture or my demeanour are among the vehicles of expression that constitute the first expressive episode of my belief that the Earth goes around the Sun: the silent judgement [The Earth goes around the Sun] with which I conclude the silent deliberation. For, in both cases, those vehicles of expression (e.g., my facial expressions, my bodily postures and my demeanours) occupy the same positions in the context of the expressive pattern of my deliberation and in the context of the expressive pattern of my newly formed belief as the relevant aloud utterances (vehicles of expression) of my ongoing deliberation (expressive episodes) and as the relevant aloud utterance (vehicle of expression) of my conclusion or silent judgement (expressive episode) would have occupied in a counterfactual situation in which I were speaking my mind rather than suppressing my utterances.

Let's move on to the epistemological aspect of the problem. If silent thoughts and silent judgements are episodes of expression constituted of unclear expressive vehicles (e.g., facial expressions, demeanours, bodily postures, etc.) and if we cannot perceive the whole expressive content of those vehicles of expression because the appropriate perceptual conditions cannot take place, how can we know the expressive content of those vehicles of expression? How can we ever know what other people are thinking? Let's focus on my thought [The Earth goes around the Sun]. At the moment that I think [The Earth goes around the Sun], the expressive content of my current expressive behaviour (i.e., my facial expression, my bodily posture, my demeanour…) includes the content of the thought [The Earth goes around the Sun]. However, insofar as the aloud utterance "The Earth goes around the Sun" (expressive vehicle) has been suppressed, the expressive content of my current expressive behaviour (i.e., my facial expression, my body posture, my demeanour, etc.) is *presented* to the perceiving subject in an *unclear way*, and so, the appropriate perceptual conditions to perceive the whole expressive content of my expressive behaviour don't take place. However, the fact that the precise expressive content of my current expressive behaviour cannot be *perceived* by a subject S at the time t_1 because the appropriate perceptual conditions don't take place doesn't mean that the expressive content of my current expressive behaviour cannot be *known* by S at all. For S could acquire at t_2 enough context to be able to know that the expressive content of my expressive behaviour at t_1 includes the content of the thought [The Earth goes around the Sun]; for instance, S could ask me at t_2 "What are you thinking about?" and I could tell her "I was thinking that the Earth goes around the Sun" (first-person or third-person memory-self-interpretation).

To better understand this idea, it might be helpful to present four different cases in which a subject manifests a set of *expressive episodes* with the same *expressive content* but in different *vehicles of expression*, being so the case that the degree of clarity with which that expressive content is presented to other people's perception differs depending on the vehicles of expression that are used. Thus, the four cases that are going to be presented in what follows must be understood as a continuum in which the vehicles of expression used and the degree of presentation-clarity of their expressive content is the main variable among them.

1) I am *silently* thinking about Moore's paradox and, among the chain of thoughts that cross my mind, the following thoughts take place: [If Moore's sentences were contradictions, it would be explained why their assertions usually strike us as obviously irrational. But it needs to be explained as well why in some contexts it is not irrational to assert a Moore's sentence]. Since I am thinking all of that while having a stroll around my hometown and I already have the stereotypical reputation of a philosopher, I dissimulate that I am reflecting on Moore's paradox while walking gently with a poker face (vehicles of expression), as if I were enjoying a relaxing stroll.

2) I am *silently* thinking about Moore's paradox and, among the chain of thoughts that cross my mind, the following thoughts take place: [If Moore's sentences were contradictions, it would be explained why their assertions usually strike us as obviously irrational. But it needs to be explained as well why in some contexts it is not irrational to assert a Moore's sentence]. Since I do not care about what my neighbours think about me, I don't dissimulate that I am reflecting on something important, so I walk with a frown, looking into the distance and rubbing my beard (vehicles of expression).

3) I am thinking about Moore's paradox and eventually, believing that I am alone, I *murmur* (vehicle of expression) to myself "If Moore's sentences were contradictions, it would be explained why their assertions usually strike us as obviously irrational. But it needs to be explained as well why in some contexts it is not irrational to assert a Moore's sentence" while I walk with a frown, looking into the distance and rubbing my beard (vehicles of expression).

4) I am walking in the company of a friend while reflecting on Moore's paradox and eventually I decide to tell him "If Moore's sentences were contradictions, it would be explained why their assertions usually strike us as obviously irrational. But it needs to be explained as well why in some contexts it is not irrational to assert a Moore's sentence".

As I understand them, the expressive content of my expressive behaviour in all four of these cases is the same (regarding my attitude to Moore's paradox and not regarding other mental states, like desires or intentions[64]): I express that I am deliberating about Moore's paradox and the content [If Moore's sentences were contradictions, it would be…] is an *aspect* of that total expressive content of my expressive behaviour (for that content is an aspect of my deliberation about Moore's paradox). The difference between the four cases is the *degree of clarity* with which the expressive content of my expressive behaviour is presented to a perceiving subject at t_1; namely, the *degree of clarity* with which my vehicles of expression present their expressive content to a perceiving subject at t_1. In the first case, 1) a random perceiving subject without context won't be able to tell whether I am enjoying a stroll or struggling with Philosophy. In the second case, 2) a random perceiving subject without context might be able to tell that I am thinking about something that is important to me (for I walk with a frown, looking into the distance and rubbing my beard), but not that I am thinking about Moore's paradox. In the third case, 3) a random perceiving subject without additional context might be able to tell that I am thinking about Moore's paradox if it happens that, unbeknownst to me, she is walking close enough to partially overhear me murmuring "If Moore's sentences were contradictions, it would be…". And, finally, 4) my friend can know without any additional context that I am thinking about Moore's paradox when she hears me saying "If Moore's sentences were contradictions, it would be…". However, the reason why the random perceiving subjects of 1) and 2) cannot know that I am thinking about Moore's paradox is not that my expressive behaviour doesn't have the same expressive content as in 3) and 4), but that in 1) and 2) the expressive content of my expressive behaviour is presented in a less clear way (i.e., in less clear vehicles of expression) so that the perceiving subjects need additional context to be able to know more about the expressive content of my expressive behaviour.

To prove this claim, let's see what would happen if the perceiving subjects 1) and 2) had more context. Firstly, imagine that in case 1) my neighbour sees me at t_1 walking in an apparently relaxed way, he asks me "Are you enjoying your stroll?", and I answer "Well, I was thinking that 'If Moore's sentences were contradictions, it would be…'". Apart from deciding not to ask me anything ever again, my neighbour would acquire from that answer at t_2 the context that he needs to be able to know more details of the expressive content of my expressive

[64] For instance, since in the fourth case I have the *additional* intention to tell my friend, my utterance "If Moore's sentences were contradictions, it would be […]" is not only an expressive episode of my deliberation about Moore's paradox, but also an expressive episode of my intention to tell my friend.

behaviour at t_1, including the content of my silent thought [If Moore's sentences were contradictions, it would be…] (for this content is an aspect of the total expressive content of my expressive behaviour at t_1). If I manage to issue at t_2 a veridical (first-person or third-person) self-interpretation of my thoughts at t_1, my neighbour could acquire actual knowledge (true warranted belief) of the thought included in the expressive content of my expressive behaviour at t_1. Secondly, imagine that in the case 2) the person who sees me walking with a frown, looking into the distance and rubbing my beard is a friend of mine instead of a random person and imagine that she knows that I just spent the morning studying and writing about Moore's paradox. Thanks to that piece of context about what I was doing before t_1, she might be able to know when she perceives my expressive behaviour at t_1, not only that I am thinking about *something* (as a random perceiving subject could know), but also that I am likely thinking about *Moore's paradox*. If she wants to confirm that or to know more about my particular chain of thoughts, she can ask me at t_2 and I could give her the veridical (first-person or third-person) self-interpretation "I was thinking that if Moore's sentences were contradictions, it would be…".

Therefore, silent thoughts (e.g., [If Moore's sentences were…]) are *expressive episodes* of mental states (e.g., of my deliberation about Moore's paradox) because thoughts are expressed in people's facial expressions, bodily postures or demeanours (vehicles of expression) when their facial expressions, bodily postures or demeanours occupy the same place in the context of the expressive pattern of a mental state (e.g., belief, desire, intention, emotion…) than the relevant pronounced utterances (other vehicles of expression) could have occupied to constitute the expressive episode in question. Moreover, we are not able to perceive the content of the silent thoughts of other people just by looking at their facial expressions, bodily postures or demeanours because the precise expressive content of those vehicles of expression is presented in a very unclear way, and so, the appropriate perceptual conditions cannot take place. However, that doesn't mean that we cannot know the precise expressive content of those vehicles of expression, for we could acquire that knowledge by having more context, by perceiving a bigger portion of the subject's expressive behaviour (e.g., asking the subject), etc.

As a result, the behavioural-expressivist account can appropriately explain Moore's paradox. For it appropriately explains why it is irrational to silently judge from the first-person perspective "p, but I don't believe that p" and "p, but I believe that not-p". To silently judge "p, but I don't believe that p" at t_1 and to assert "p, but I don't believe that p" at t_2 are two *numerically different* episodes of expression, but they are *typologically* identical episodes of expression (i.e., they are the same kind of expressive episode: they are acts of assent to the content [p, but I don't believe that p]). They are one and the same kind of *expressive episode* instantiated two different times (at t_1 and t_2) in two different sets of *expressive vehicles*: the silent judgement (i.e., act of assent to [p, but I don't believe that p]) is instantiated *only* in facial expressions, bodily postures or demeanours (for the aloud utterance has been supressed), while the assertion (i.e., act of assent to [p, but I don't believe that p]) is instantiated *also* in the aloud utterance "p, but I don't believe that p". And the same goes for the Moore's sentence "p, but I believe that not-p". Therefore, the expressive content of silently judging "p, but I don't believe that p" and "p, but I believe that not-p" from the first-person perspective is as self-contradictory-like and as self-contradictory as the expressive content of asserting "p, but I don't believe that p" and "p, but I believe that not-p" from the first-person perspective: they are the same kind of expressive episodes instantiated in two different sets of expressive vehicles.

Conclusion

In this essay, it has been argued that the behavioural-expressivist account of Transparency of belief, together with the non-relational view of expression and the expressivist view of first-person self-knowledge that follows from it, helps to understand the phenomena of self-deception and Moore's paradox in the appropriate way (as it is proved by the fact that it is able to deliver the best account of Moore's paradox and the best account of self-deception currently available in the literature).

The phenomenon of *Transparency of belief* consists in the fact that the question "Do you believe that p?" is answered in the same way as the question "Is p the case?" when it is answered from the first-person deliberative perspective: deliberating about whether p on the basis of evidence about p until we make up our own minds. The phenomenon of *self-deception* consists in the fact that sometimes subjects sincerely avow "I believe that p" from the first-person perspective while acting in a conflicting way with the mental state that they make explicit in the avowal (which shows that self-deceivers suffer from some kind of lack of self-knowledge). And the phenomenon of *Moore's paradox* consists in the fact that sentences like "p, but I don't believe that p" or "p, but I believe that not-p" are sometimes irrational to assent in spite of the fact that they can be true$_{rs}$. What these phenomena have in common is that it is usually considered in the literature that they are the result of the success (i.e., Transparency) or failure (i.e., self-deception and Moore's paradox) of the special first-person procedure of belief-formation responsible for the *groundless* (i.e., made on the basis of no evidence about the subject's mental states) and *authoritative* (i.e., presumably true) character of first-person avowals (e.g., "I believe that p"), and so, the result of the success or failure of the first-person procedure of belief-formation responsible for epistemic and authoritative self-knowledge (i.e., strongly warranted true$_{rs}$ beliefs).

Epistemic accounts of Transparency consider that the transparent question "Do you believe that p?" asks about the subject's beliefs and that it is answered in the same way as the question "Is p the case?" because first-person avowals are groundless and authoritative self-ascriptions of mental states made on the basis of the special first-person procedure responsible for authoritative and epistemic self-knowledge: a procedure that, regardless of the particularities of each account, always involves a first-person deliberation about whether p. *Neo-expressivist accounts* have in common with epistemic accounts of Transparency the following three ideas. Firstly, the idea that first-person avowals are groundless and authoritative *self-ascriptions* of mental states (for neo-expressivist accounts consider that first-person avowals express the same mental state that they self-ascribe). Secondly, the idea that first-person self-knowledge is an *epistemic phenomenon* that consists in having a true$_{rs}$ belief about one's own mental states warranted in a stronger type of way thanks to a special first-person procedure to access one's own mental states (which could be based on deliberation about whether p —epistemic accounts of Transparency— or in the expressive mechanism responsible for first-person avowals —neo-expressivist accounts—). And thirdly, the idea that mental states and their characteristic set of expressions are two different items related in some way (i.e., *relational view of expression*) so that when I acquire first-person authoritative and epistemic self-knowledge, I have exclusive access (thanks to the true$_{rs}$ strongly warranted belief about my own mental states) to a mental item of mine different from expression that cannot be accessed by other people.

In regard to self-deception, there are four different kinds of accounts. *Intentionalist accounts* consider that self-deceivers' inconsistent behaviour is explained because self-deceivers have and fulfil the intention to deceive themselves by believing something that they consider to be false. *Motivationalist accounts* consider that self-deceivers' inconsistent behaviour is explained because self-deceivers have a motivated bias when deliberating about whether p to form the belief that p. *Epistemic accounts* consider that the self-deceivers' inconsistent behaviour is the result of an epistemic failure in the special first-person process responsible for Transparency and for first-person epistemic and authoritative self-knowledge. And, finally, *psychological-state accounts* of self-deception consider that the self-deceivers' inconsistent behaviour is the result of a *sui generis* mental state different from belief and from any other mental state. All these accounts of self-deception are compatible with the three ideas shared by epistemic accounts of Transparency and by neo-expressivist accounts. Furthermore, epistemic accounts of self-deception explicitly endorse those three ideas (at least in the case of Fernández's account). Again, these ideas are the following: the idea that first-person avowals are groundless and authoritative self-ascriptions of mental states, the idea that there is a special first-person procedure of belief-formation responsible for first-person epistemic and authoritative self-knowledge (i.e., true$_{rs}$ strongly warranted belief), and the idea that mental states and their characteristic set of expressions are two different items related in some way (i.e., relational view of expression).

In regard to Moore's paradox, there are four different kinds of accounts. All these accounts of Moore's paradox (except for Linville's & Ring's account) have in common the idea that the first-person avowals "I don't believe that p" and "I believe that not-p" are self-ascriptions of beliefs rather than judgements about whether p. *Pragmatic accounts* consider that Moore's sentences (i.e., "p, but I don't believe that p" and "p, but I believe that not-p") are irrational to assent in spite of the fact that they can be true$_{rs}$ because they don't have appropriate conditions of assertion. *Psychological accounts* consider that Moore's sentences are irrational to assent to in spite of the fact that they can be true$_{rs}$ because they involve some kind of inconsistency among the subject's mental states (i.e., among the subject's conscious beliefs or among the subject's commitments). *Epistemic accounts* consider that Moore's sentences are irrational to assent to in spite of the fact that they can be true$_{rs}$ because they are the result of an epistemic failure in the first-person procedure of belief-formation responsible for Transparency and for first-person epistemic and authoritative self-knowledge. And *semantic accounts* consider that Moore's sentences are irrational to assent to because they are self-contradictions. All these accounts of Moore's paradox are compatible with the three ideas shared by epistemic accounts of Transparency and by neo-expressivist accounts. Furthermore, epistemic accounts of Moore's paradox explicitly endorse them. Again, these ideas are the following three: the idea that first-person avowals are groundless and authoritative self-ascriptions of mental states, the idea that there is a special first-person procedure of belief-formation responsible for first-person epistemic and authoritative self-knowledge (i.e., true$_{rs}$ strongly warranted belief), and the idea that mental states and their characteristic set of expressions are two different items related in some way (i.e., relational view of expression).

As it was said before, the main claim argued in this essay is that the behavioural-expressivist account of Transparency explains self-deception and Moore's paradox better than epistemic accounts of Transparency. And the auxiliary claim argued in this essay is that the behavioural-expressivist account of Transparency is able to do so thanks to the fact that it endorses a non-relational view of expression and a non-epistemic view of first-person self-knowledge.

The behavioural-expressivist account of Transparency endorses a semantic view of Transparency according to which Transparency is explained in the following way. The question "Do you believe that p?" can be meant either in a deliberative or in a self-ascriptive way. When it is meant in a deliberative way, it is transparent to "Is p the case?" because both questions have the same meaning, and so, the answer to that question is supposed to be an expressive episode of attitude consisting in a *judgement about whether p* (i.e., an expressive episode of belief that can take the linguistic form of an assertion —e.g., "p is the case"— or the linguistic form of a first-person avowal —e.g., "I believe that p"—) made by deliberation on the basis of evidence about whether p. The subject will have *first-person expressive self-knowledge* at answering this question if she has the ability to appropriately express that attitude (i.e., in the appropriate circumstances, using the appropriate vehicle of expression, etc.) and if the episode of attitude is expressed in a self-conscious way (i.e., knowing what one is up to). By contrast, when the question "Do you believe that p?" is meant in a self-ascriptive way, it is not transparent to the question "Is p the case?" because it is meant as a question about the subject's beliefs, and so, the answer to that question is supposed to be an episode of expression of a second-order belief consisting in a *self-ascription of attitude* (i.e., an expressive episode of the second-order belief that I believe that p that can take the linguistic form of an assertion —e.g., "It is the case that I believe that p"— or the linguistic form of a third-person avowal —e.g., "I believe that p"—) made by self-inspection on the basis of evidence about whether one believes that p. The subject will have *third-person epistemic self-knowledge* if that second-order belief is warranted and true, and she might even have third-person authoritative self-knowledge if she manages to warrant that second-order belief to a higher degree than the beliefs of other subjects about her own mental states can possibly be using memory and/or introspection as epistemic sources of evidence about her own mental states (for, unlike perception and inference, memory and introspection are sources of evidence exclusive of self-inspection). Since epistemic self-knowledge is a third-person self-inspective phenomenon, the idea of exclusive first-person access to a mental item of mine is ruled out from the outset, and so, it follows a non-relational view of expression and mental states: mental states are nothing over and above patterns of expressive behaviour that might be publicly available to other people's perception (assuming that the appropriate perceptual conditions take place).

From the behavioural-expressivist account of Transparency, together with the auxiliary claim of the non-relational view of expression and the auxiliary claim of the first-person expressive self-knowledge and third-person epistemic self-knowledge, the following accounts of self-deception and Moore's paradox have been given.

In regard to self-deception, behavioural expressivism explains self-deception claiming that it is a *sui generis* mental state that belongs to the class of unconscious mental states and that has been formed by a motivated bias in the first-person deliberation about whether p that causes the following expressive failure: the subject makes a judgement about whether p, but she ends up forming an unconscious attitude (i.e., self-deception) rather than a conscious attitude (e.g., belief) because of the motivated bias. Since self-deception is an unconscious mental state, it involves lack of first-person expressive self-knowledge and difficulties in acquiring third-person epistemic self-knowledge. On the one hand, self-deceived subjects don't have first-person expressive self-knowledge because when they answer the transparent question "Do you believe that p?" from the first-person deliberative perspective, they answer with a judgement about whether p that is a non-self-conscious episode of expression (i.e., they don't know what they are up to): they think that they are exercising the ability to express a conscious attitude (e.g., belief), but they are actually exercising the ability to express an unconscious attitude (i.e., their self-deceived mental state about p). Since this occurs each time that self-deceived subjects express their self-deceived mental state from the first-person deliberative perspective (for self-deceived mental states are unconscious mental states and unconscious mental states cannot be self-consciously expressed), self-deceived subjects lack first-person expressive self-knowledge of their self-deceived mental state. On the other hand, when self-deceived subjects answer the non-transparent question "Do you believe that p?" from the third-person self-inspective perspective, they answer with a self-ascription of attitude that has a low chance of being true, and so, a low chance of being an instance of third-person epistemic self-knowledge (i.e., true warranted belief). The reasons why self-deceived subjects have a low chance of acquiring third-person epistemic self-knowledge of their self-deceived mental state are two. Firstly, the episodes characteristic of the expressive pattern of self-deceived mental states are similar in appearance to the episodes characteristic of the expressive pattern of some other conscious mental attitudes (e.g., belief). And secondly, self-deceived subjects are motivationally biased to gather and assess the evidence about which attitude they have about p in a biased way because of the same motivational state that biased their deliberation about whether p in the first place (e.g., the desire to be healthy, anxiety about the possibility of being ill, etc.).

In regard to Moore's paradox, it has been argued that the behavioural-expressivist account of Transparency explains Moore's paradox in the following way. On the one hand, when "I don't believe that p" or "I believe that not-p" are issued from the third-person self-inspective perspective, as if they were answers to the self-ascriptive question "Do you believe that p?", it is not irrational to assent to the Moore's sentences "p, but I don't believe that p" or "p, but I believe that not-p" because they are not self-contradictions-like nor self-contradictions: "p" assents to the fact that p and expresses the subject's belief that p, "I don't believe that p" is a self-ascription of lack of belief about p and expresses the subject's second-order belief that she lacks a belief about p, and "I believe that not-p" is a self-ascription of the belief that not-p and expresses the subject's second-order belief that she has the belief that not-p. Then, in these conversational contexts, Moore's sentences can be both true$_{nrs}$ (their expressive content is a possible combination of mental states: the judgement that p and the second-order belief that one doesn't have any belief about p or the second-order belief that one has the belief that not-p) and true$_{rs}$ (their intentionality or propositional content is a possible state of affairs). On the other hand, when "I don't believe that p" or "I believe that not-p" are issued from the first-person deliberative perspective, as if they were answers to the deliberative question "Do you believe that p?", it is irrational to assent to the Moore's sentences "p, but I don't believe that p" or "p, but I believe that not-p" because they are self-contradictions-like, and so, self-contradictions. They are self-contradictions-like because they *apparently* express two incompatible attitudes (i.e., the belief that p and lack of belief about p, or the belief that p and the belief that not-p) so that they cannot have possible truth$_{nrs}$-conditions: they don't have any expressive content because they don't express any mental state. As a result, they are self-contradictions as well because they are *nonsenses*: since such acts of assent don't have any expressive content, they don't have any intentionality or propositional content either, and so, they don't say anything about the world (neither true$_{rs}$ nor false$_{rs}$). Therefore, they are self-contradictions because, insofar as they are nonsenses, they don't have possible truth$_{nrs}$-conditions.

Furthermore, in addition to the main claim (i.e., that the behavioural-expressivist account of Transparency, together with the non-relational view of expression and the expressivist view of first-person self-knowledge, explains self-deception and Moore's paradox), three auxiliary arguments have been given in favour of the behavioural-expressivist interpretation of Transparency, in favour of the non-relational view of expression, and in favour of the behavioural-expressivist view of third-person epistemic self-knowledge. Firstly, it has been argued that the behavioural-expressivist account of Transparency explains the phenomenon of Transparency better than epistemic accounts of Transparency for two reasons. On the one hand, there are examples of attitudes that are transparent to the world and of which the subject has first-person self-knowledge without the need to deliberate again about whether p: attitudes that are already held by the subject. For instance, if you ask me "Do you believe that the Earth goes around the Sun?", I can give you an answer from the first-person deliberative perspective without deliberating about the issue again, just as I could do if you would ask me "Does the Earth go around the Sun?". Therefore, it seems that the question "Do you believe that the Earth goes around the Sun?" is transparent to the question "Does the Earth go around the sun?" in this case, and that I have first-person self-knowledge of the answer without the need to deliberate about whether the Earth goes around the Sun again. Epistemic accounts of Transparency have problems explaining this type of cases because they claim that Transparency is the result of a special first-person procedure, responsible for authoritative and epistemic first-person self-knowledge, that is triggered by deliberation about whether p. Then, if no deliberation occurs again when answering the transparent question "Do you believe that p?" in the case of attitudes that are already held by the subject, it seems that they can't explain why subjects have first-person self-knowledge to answer the question and why Transparency still occurs. By contrast, the behavioural-expressivist account of Transparency doesn't have difficulties in explaining this type of cases. I can answer the transparent question "Do you believe that the Earth goes around the Sun?" from the first-person deliberative perspective without deliberating about the issue again because I already formed the belief that the Earth goes around the Sun through first-person deliberation and now I can express a new episode of that belief on the basis of no extra evidence. Also, I have first-person expressive self-knowledge because I have the ability to express my belief in an appropriate way and because I answer the question self-consciously. On the other hand, the behavioural-expressivist account of Transparency explains the deliberation about whether p characteristic of the phenomenon of Transparency in a more plausible way than epistemic accounts of Transparency. For this deliberation is about whether p and the behavioural-expressivist account claims that such deliberation about whether p concludes with a judgement about whether p (e.g., "p is the case" or "I believe that p") that is an expressive episode of the newly formed attitude (e.g., belief), whereas epistemic accounts of Transparency claim that such deliberation about whether p concludes with a self-ascription of attitude rather than with a judgement about whether p.

Secondly, it has been argued that the non-relational view of expression that follows from the behavioural-expressivist account of Transparency has different advantages over the relational view of expression endorsed by epistemic accounts of Transparency and by neo-expressivist accounts. The first reason that has been given in favour of the non-relational view of expression goes as follows. Relational views of expression follow from the idea that expressions are symptomatic evidence of mental states (i.e., causal relational views) or from the idea that expressions are defeasible criterial evidence of mental states (i.e., constitutive or mereological relational views), whereas the non-relational view of expression follows from the idea that expressions are indefeasible criterial evidence of mental states. Since understanding criteria as indefeasible evidence seems to be the best way to explain how evidence can warrant beliefs to produce knowledge (i.e., true warranted belief) because indefeasible criteria guarantees the occurrence of that of which something is criterial evidence of, the non-relational view of expression has the advantage of being the only view of expression that follows from and that is compatible with the idea that expressions are indefeasible criteria of mental states. The second argument that has been offered in favour of the non-relational view of expression endorsed by behavioural expressivism is that the non-relational view of expression better explains the phenomenon of pretence and dissimulation than relational views of expression because it is able to explain cases of pretence and dissimulation and cases of discovering that someone is pretending or dissimulating with the same theoretical resources. According to behavioural expressivism, the explanation goes as follows. On the one hand, pretending that one has M is all about manifesting an expressive pattern similar in appearance to the expressive pattern of M, and discovering that someone else is pretending to have M is all about discovering the differences between the real expressive pattern of M and the apparently similar expressive pattern of pretending M. On the other hand, dissimulating that one has M is all about repressing the most intense episodes of M, and discovering that someone else is dissimulating M is all about discovering the details of her expressive behaviour that reveals that she actually has M. By contrast, relational views of expression consider that pretending M is manifesting the expressive episodes characteristic of M without having M, and that dissimulating M is suppressing all the expressive episodes of M while having M (or, at least, they think that this is a conceptually possible case of dissimulation). As a result, relational views of expression need to explain in addition how it is that people are sometimes able to bypass the expressions of a subject to find out whether she is pretending M (i.e., she doesn’t have M in spite of expressing M) or dissimulating M (i.e., she does have M in spite of not expressing M at all).

Thirdly, it has been argued that the view of self-knowledge that follows from the behavioural-expressivist account of Transparency explains both the phenomenon of first-person self-knowledge as expressive self-knowledge and the phenomenon of third-person self-knowledge as epistemic self-knowledge better than other competing accounts. Particularly, better than the epistemic view of first-person self-knowledge endorsed both by epistemic accounts of Transparency and by neo-expressivist accounts and better than the inferential view of third-person epistemic self-knowledge defended by Cassam (2014). Regarding first-person epistemic self-knowledge, it has been argued that the idea of first-person epistemic self-knowledge is conceptually flawed. The idea of first-person epistemic self-knowledge involves the idea of exclusive first-person access, the idea of access involves the idea of an ontologically robust item (i.e., possibly accessed on different occasions), but the idea of first-person exclusive access is incompatible with the idea of access to an ontologically robust item because there is no standard of accuracy other than what the subject claims to have accessed at each particular time (i.e., there is no *independent* standard of accuracy, which is the role that an ontologically robust item is supposed to play). Then, the idea of first-person self-knowledge as an expressive phenomenon arises as a viable alternative. Subjects might have first-person self-knowledge of a mental state both in the sense that they might have the ability to express that mental state in an appropriate way (e.g., in the appropriate situations, in the appropriate vehicle of expression, etc.) and in the sense that they might express that mental state self-consciously (i.e., knowing what they are doing).

In regard to third-person epistemic self-knowledge, it was argued that Cassam's inferential account wasn't able to explain third-person epistemic self-knowledge because the evidence on the basis of which an inference is made eventually stops being inferential (e.g., sometimes a feeling cannot be inferred from anything else), and so, Cassam's account doesn't explain how subjects could use that evidence to infer their own mental states (remember Katherine's example). Then, once again, the alternative is to understand third-person epistemic self-knowledge in a behavioural-expressivist way. Third-person epistemic self-knowledge is the result of a third-person process of self-inspection that consists in making a judgement about one's own mental states on the basis of evidence provided by four possible epistemic sources: perception (e.g., seeing my tired face in the mirror), inference (e.g., that my friends tell me that I look tired), memory (e.g., that I remember feeling tired —even if it is just a second ago—) and introspection (e.g., that I would feel happy when I finish my PhD). Since memory and introspection present me as *expressing* an episode of mental state rather than as perceiving or inferring one of my expressive episodes, memory and introspection are responsible for the fact that self-inspection sometimes can deliver third-person authoritative self-knowledge when they are used as epistemic sources of evidence. Indeed, since memory presents me as expressing an episode of mental state in the past (e.g., as feeling tired —even if it was just a second ago—) and introspection presents me as expressing an episode of mental state in a hypothetical situation (e.g., as feeling happy when I finish my PhD), I can warrant (in the absence of any setback) my self-inspective second-order beliefs about my own mental states on the basis of evidence provided by memory and introspection better (in degree) than other people can warrant their beliefs about my mental states on the basis of evidence provided by perception and inference. Moreover, the fact that memory and introspection are sources of evidence that present me as expressing my episodes of mental states (rather than as perceiving or inferring them) explains how subjects can have third-person (authoritative) self-knowledge of their own feelings, avoiding so the problem raised against Cassam's inferential account.

Therefore, it can be concluded that behavioural expressivism is the right account of the phenomenon of Transparency, of the nature of mental states and of first-person and third-person self-knowledge. Transparency is an epistemic phenomenon that arises because of the fact that the difference between the first-person deliberative perspective and the third-person self-inspective perspective is semantic rather than epistemic: whether one makes a judgement about whether p on the basis of evidence about p (first-person deliberative perspective) or whether one makes a judgement about one's own mental states on the basis of evidence about one's own mental states (third-person self-inspective perspective). Mental states are temporal patterns of expressive behaviour that are publicly manifested each time that a subject manifests a particular expressive episode of a particular mental state. First-person self-knowledge is an expressive phenomenon that has to do with two expressive properties of our mental states: whether the subject has the ability to express a particular mental state in an appropriate way and whether a particular mental state is self-consciously expressed. And third-person self-knowledge is an epistemic phenomenon that has to do with making self-inspective judgments about the mental states that one has on the basis of the evidence about one's own mental states provided by memory and introspection, and exceptionally, by perception and inference as well. When the judgement is made mainly on the basis of the former two, third-person authoritative self-knowledge can possibly occur.

References

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

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington D.C., American Psychiatric Association.

Armstrong, D. (1995). “The causal Theory of the Mind”. Lyons, W. (ed.), *Modern Philosophy of Mind*, London, Everyman, 175-190.

Audi, R. (1982). “Self-Deception, Action, and Will”. *Erkenntnis*, 18 (2), 133-158.

Audi, R. (1988). “Self-deception, rationalization, and reasons for acting”. McLaughlin, B. P. & Rorty, A. O. (eds.), *Perspectives on Self-Deception*, Berkeley and Los Angeles, University of California Press, 92-120.

Audi, R. (1989), “Self-Deception and Practical Reasoning”, *Canadian Journal of Philosophy*, 19 (2), 247-266.

Audi, R. (1997). *Moral Knowledge and Ethical Character*, New York, Oxford University Press.

Austin, J. L. (1962). *How to Do Things with Words*, Oxford, Oxford University Press.

Austin, J. L. (1970). “Performative utterances”. *Philosophical Papers*, Oxford, Oxford University Press, 233-252.

Ayer, A. J. (2001). *Language, truth and logic*, London, Penguin.

Bach, K. (1981). “An Analysis of Self-Deception”. *Philosophy and Phenomenological Research*, 41 (3), 351-370

Baldwin, T. (1990). *G.E. Moore*, London and New York, Routledge.

Baldwin, T. (2007). “The normative character of belief”. Green, M. S. & Williams, J. N. (eds.), *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*, Oxford, Oxford University Press, 76-89.

Barnes, A. (1997). *Seeing through Self-deception*, Cambridge, Cambridge University Press.

Bar-On, D. & Sias, J. (2013). “Varieties of expressivism”. *Philosophy Compass*, 8 (8), 699-713.

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

Bayne, T. & Fernandez, J. (eds.) (2008). *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, New York and Hove, Psychology Press.

Beattie, J. & Baron, J. (1988). “Confirmation and matching biases in hypothesis testing”. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 40 (2), 269-297.

Bermúdez, J. (1997). “Defending intentionalist accounts of self-deception”. *Behavioral and Brain Sciences*, 20 (1), 107-108.

Bermúdez, J. (2000). “Self-deception, intentions and contradictory beliefs”. *Analysis*, 60 (4), 309-319.

Boyle, M. (2009). “Two kinds of self-knowledge”. *Philosophy and Phenomenological Research,* 78 (1), 133-164.

Boyle, M. (2011). “Transparent self-knowledge”. *Aristotelian Society: Supplementary Volume*, 85 (1), 223-241.

Boyle, M. (2015). “Critical Study: Cassam on Self-Knowledge for Humans”. *European Journal of Philosophy*, 23 (2), 337-348.

Byrne, A. (2005). “Introspection”. *Philosophical Topics*, *33* (1), 79-104.

Byrne, A. (2011). "Transparency, belief, intention". *Aristotelian Society Supplementary Volume*, 85 (1), 201–221.

Byrne, A. (2018). *Transparency and Self-Knowledge*, Oxford, Oxford University Press.

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

Cassam, Q. (2014). *Self-Knowledge for Humans*, New York, Oxford University Press.

Chrisman, M. (2009). "Expressivism, truth, and (self-) knowledge". *Philosophers' Imprint*, 9 (3), 1-26.

Coliva, A. (2016). *The Varieties of Self-knowledge*, London, Palgrave Macmillan.

Davidson, D. (2004). *Problems of Rationality*, Oxford, Oxford University Press.

Demos, R. (1960). "Lying to oneself". *Journal of Philosophy*, 57 (18), 588-595.

Edgley, R. (1969). *Reason in Theory and Practice*, London, Hutchinson.

Egan, A. (2008). "Imagination, delusion, and self-deception". Bayne, T. & Fernandez, J. (eds.) (2010). *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, New York and Hove, Psychology Press, 263-277.

Evans, G. (1982). *The Varieties of Reference*, New York, Oxford University Press.

Fernández, J. (2005). "Self-Knowledge, Rationality and Moore's Paradox". *Philosophy and Phenomenological Research*, 71 (3), 533-556.

Fernández, J. (2012). "Self-deception and self-knowledge". *Philosophical Studies*, 162 (2): 379–400.

Fernández, J. (2013). *Transparent minds*, Oxford, Oxford University Press.

Finkelstein, D. (1999). "On the distinction between conscious and unconscious states of mind". *American Philosophical Quarterly,* 36 (2), 79-100.

Finkelstein, D. (2003). *Expression and the Inner*, Cambridge (Massachusetts) and London (England), Harvard University Press.

Foss, J. (1980). "Rethinking Self-Deception". *American Philosophical Quarterly*, 17 (3), 237–243.

Funkhouser, E. & Barrett, D. (2016). “Robust, unconscious self-deception: Strategic and flexible”. *Philosophical Psychology*, 29 (5), 1-15.

Funkhouser, E. (2005). “Do the Self-Deceived Get What They Want?”, *Pacific Philosophical Quarterly*, 86(3), 295–312.

Gallois, A. (1996). *The World Without, the Mind Within: An Essay on First-Person Authority*, Melbourne, Cambridge University Press.

García Rodriguez, Á. (2018). “Direct Perceptual Access to Other Minds”. *International Journal of Philosophical Studies*, 26 (1), 24-39.

García Rodriguez, Á. (2019a). “Expression and the transparency of belief”. *Eur J Philos*, 27 (1), 136-147.

García Rodriguez, Á. (2019b). “A Wittgensteinian View of Mind and Self-Knowledge”. *Philosophia*, DOI: 10.1007/s11406-019-00143-y.

Gaynesford, R. M. (2002). “Blue book ways of telling: Criteria, openness and other minds”. *Philosophical Investigations*, 25 (4), 319–330.

Gendler, T. (2010). *Intuition, Imagination, and Philosophical Methodology*, Oxford, Oxford University Press.

Gertler, B. (2010). *Self-Knowledge*, New York, Routledge.

Goldie, P. (2011). “Grief: A narrative account”. *Ratio*, 24 (2), 119-137.

Hamilton, A. (2014). *Routledge Philosophy Guidebook to Wittgenstein and on Certainty*, London and New York, Routledge.

Heal, J. (1994). “Moore's paradox: A Wittgensteinian approach”. *Mind*, 103 (409), 5-24.

Hempel, C. (1980). “The logical analysis of psychology”. In Ned Block (ed.). *Readings in Philosophy of Psychology*, Cambridge, Harvard University Press, 14-23.

Hirstein, W. (2005). *Brain Fiction: Self-Deception and the Riddle of Confabulation*, Cambridge (Massachusetts) and London (England), MIT Press.

Holton, R. (2001). “What is the role of the self in self-deception?”. *Proceedings of the Aristotelian Society*, 101 (1), 53-69.

Johnston, M. (1988). “Self-Deception and the Nature of Mind”. McLaughlin, B.P. & Rorty, A. O. (eds.). *Perspectives on Self-Deception*, Berkeley and Los Angeles, University of California Press, 63-91.

Keeling, S. (2018). “Confabulation and rational obligations for self-knowledge”. *Philosophical Psychology*, 31 (8), 1215-1238.

Lauria, F.; Preissmann, D. & Clément, F. (2016). “Self-Deception as Affective Coping. An Empirical Perspective on Philosophical Issues”. *Consciousness and Cognition*, 41 (1), 119-134.

Lazar, A, (1999). “Deceiving oneself or self-deceived? On the formation of beliefs ‘under the influence’”. *Mind*, 108 (430), 265-290.

Levy, N. (2004). “Self-deception and moral responsibility”. *Ratio*, 17 (3), 294-311.

Linville, K. & Ring, M. (1991). “Moore's paradox revisited”. *Synthese*, 87 (2), 295-309.

Lycan, W. G. (1971). “Noninductive Evidence: Recent Work on Wittgenstein’s ‘Criteria’”. *American Philosophical Quarterly*, 8 (2), 109-125.

Lynch, K. (2012). “On the ‘tension’ inherent in self-deception”. *Philosophical Psychology*, 25 (3), 433-450.

Malcolm, N. (1954). “Wittgenstein's philosophical investigations”. *Philosophical Review*, 63 (4), 530-59.

McDowell, J. (1998). “Criteria, Defeasibility, and Knowledge”. *Meaning, Knowledge and Reality*, Cambridge (Massachusetts) and London (England), Harvard University Press.

McHugh, C. (2013). “Epistemic responsibility and doxastic agency”. *Philosophical Issues*, 23 (1), 132-157.

Mele, A. (1999). “Twisted self-deception”. *Philosophical Psychology*, 12 (2), 117-137.

Mele, A. (2001). *Self-Deception Unmasked*, Princeton, Princeton University Press.

Mele, A. (2008). “Self-Deception and Delusions”. Bayne, T. & Fernandez, J. (eds.) (2010). *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, New York and Hove, Psychology Press, 263-277.

Moore, G. E. (1993). "Moore's Paradox". Baldwin, Thomas (ed.). *G. E. Moore: Selected Writings*, London, Routledge, 207–212

Moran, R. A. (1997). "Self-knowledge: Discovery, resolution, and undoing". *European Journal of Philosophy*, 5 (2), 141-61.

Moran, R. A. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton, Princeton University Press.

Moran, R. A. (2003). "Responses to O'Brien and Shoemaker". *European Journal of Philosophy*, 11 (3), 402-19.

Nelkin, D. K. (2002). "Self-Deception, Motivation and the Desire to Believe", *Pacific Philosophical Quarterly*, 83 (4), 384–406.

Nelkin, D. K. (2012). "Responsibility and Self-Deception: A Framework". *Humana Mente*, 5 (20), 117-139.

Patten, D. (2003). "How do we deceive ourselves?". *Philosophical Psychology*, 16 (2), 229-247.

Pears, D. F. (1984). *Motivated Irrationality*, Oxford: Oxford University Press.

Pears, D. F. (1991). "Self-deceptive belief-formation". *Synthese*, 89 (3), 393-405.

Putnam, Hilary (ed.) (1979). *Philosophical Papers: Volume 2, Mind, Language and Reality*, Cambridge, Cambridge University Press.

Rey, G. (1988). "Toward a computational account of Akrasia and self-deception". McLaughlin, B.P. & Rorty, A. O. (eds.). *Perspectives on Self-Deception*, Berkeley and Los Angeles, University of California Press, 264-296.

Rorty, A. O. (1988). "The deceptive self: Liars, layers, and lairs". McLaughlin, B. P. & Rorty, A. O. (eds.). *Perspectives on Self-Deception*, Berkeley and Los Angeles, University of California Press, 11-28.

Rosenthal, D. (2005). *Consciousness and Mind*, Oxford, Oxford University Press.

Ryle, G. (1949). *The Concept of Mind*, London, Hutchinson & Co.

Sanford, D. (1988) "Self-Deception as Rationalization". McLaughlin, B.P. & Rorty, A. O. (eds.). *Perspectives on Self-Deception*, Berkeley and Los Angeles, University of California Press, 157-169.

Scott-Kakures, D. (1997). "Self-knowledge, akrasia, and self-criticism". *Philosophia*, 25 (1-4), 267-295.

Scott-Kakures, D. (2001). "High anxiety: Barnes on what moves the unwelcome believer", *Philosophical Psychology*, 14 (3), 313-326.

Scott-Kakures, D. (2002). "At 'permanent risk': Reasoning and self-knowledge in self-deception". *Philosophy and Phenomenological Research*, 65 (3), 576-603.

Scott-Kakures, Dion (1996). "Self-deception and internal irrationality". *Philosophy and Phenomenological Research*, 56 (1), 31-56.

Shah, N. & David Velleman, J. (2005). "Doxastic deliberation". *Philosophical Review*, 114 (4), 497-534.

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

Shoemaker, S. (ed.) (1963). *Self-Knowledge and Self-Identity*, Cornell, Cornell University Press.

Silva, J., Ferrari, M., Leong, G., & Penny, G. (1998). "The dangerousness of persons with delusional jealousy". *Journal of the American Academy of Psychiatry and the Law*, 26 (4), 607–623.

Sorensen, R. (1985). "Self-Deception and Scattered Events". *Mind*, 94 (373): 64–69.

Soyka, M. (1995). Prevalence of delusional jealousy in schizophrenia. *Psychopathology*, 28 (2), 118-120.

Soyka, M., & Schmidt, P. (2011). Prevalence of delusional jealousy in psychiatric disorders. *Journal of forensic sciences*, *56* (2), 450-452.

Stevenson, C. L. (1963). *Facts and Values*. New Heaven and London, Yale University Press.

Stone, T. & Young, A. W. (1997). "Delusions and brain injury: The philosophy and psychology of belief". *Mind and Language*,12 (3-4), 327-364.

Strawson, P. F. (1954). Wittgenstein, L. *Philosophical Investigations* (Book Review). *Mind*, 63 (249), 70-99.

Sullivan-Bissett, E. (2014). "Implicit bias, confabulation, and epistemic innocence". *Consciousness and Cognition*, 33, 548-560.

Szabados, B. (1973). "Wishful thinking and self-deception". *Analysis*, 33 (6), 201-205.

Szabados, B. (1974). "Self-deception". *Canadian Journal of Philosophy*, 4 (1), 51-68

Szabados, B. (1985). "The Self, Its Passions and Self-Deception". Martin, M. W. (ed.). *Self-deception and Self-Understanding*, Lawrence, Kansas University Press.

Talbott, W. J. (1995). "Intentional self-deception in a single coherent self". *Philosophy and Phenomenological Research*, 55 (1), 27-74.

Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*, New York, Basic Books.

Trope, Y.; Gervey B. & Liberman N. (1997). "Wishful Thinking from a pragmatic hypothesis-testing perspective". Myslobodsky, M. S. (ed.). *The mythomanias: The nature of deception and self-deception*, Mahwah, Psychology Press.

Tumulty, M. (2012). "Delusions and not-quite-beliefs". *Neuroethics*, 5(1), 29-37.

Van Leeuwen, N. (2007a). "The product of self-deception". *Erkenntnis*, 67 (3), 419 - 437.

Van Leeuwen, N. (2007b). "The spandrels of self-deception: Prospects for a biological theory of a mental phenomenon". *Philosophical Psychology*, 20 (3), 329 – 348.

Williams, J. N. (2006). "Moore's paradoxes and conscious belief". *Philosophical Studies*, *127* (3), 383-414.

Witherspoon, E. (2011). "Wittgenstein on Criteria and The Problem of Other Minds". *The Oxford Handbook of Wittgenstein*. Kuusela, O. & M. McGinn (eds.). Oxford, Oxford University Press.

Wittgenstein, L. (1953). *Philosophical Investigation*. Tr. G. E. M. Anscombe. 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.