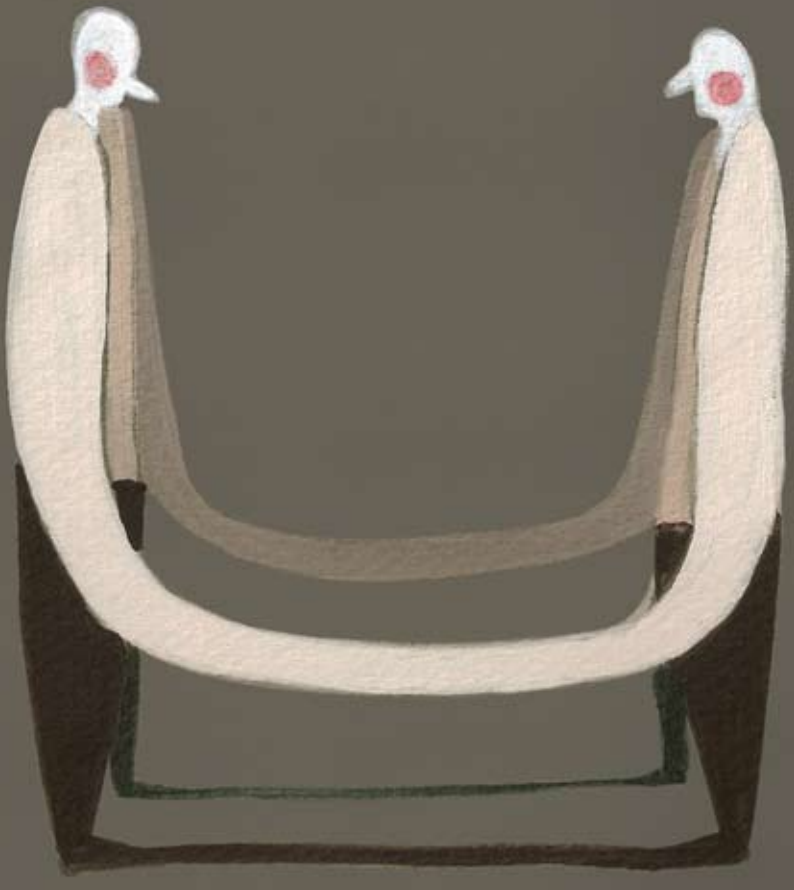
Knowing the nature of one's mind

an externalist basis for self-knowledge



Cristina Borgoni Gonçalves

tesis doctoral

Knowing the nature of one's mind

an externalist basis for self-knowledge

Tesis doctoral

Doctoranda
Cristina Borgoni Gonçalves

Director
Manuel de Pinedo García

Departamento de Filosofía I
Universidad de Granada

Granada, Octubre de 2009

Editor: Editorial de la Universidad de Granada
Autor: Cristina Borgoni Gonçalves
D.L.: GR 3862-2009
ISBN: 978-84-692-7850-5

Ilustración de la cubierta ©
Beatriz Martín

Diseño y maquetación
César Marini

CONTENTS

AGRADECIMIENTOS

En general, prefiero que un texto esté escrito en primera persona del singular. Me gusta saber que hay alguien que se arriesga con lo que está diciendo, que se expone y que asume como suya la responsabilidad de lo que se dice. En esta tesis, he intentado ejercer esta preferencia. Pero, es posible que haya cometido una serie de injusticias debido al exceso de uso del 'yo', ya que esta tesis no podría, literalmente, estar escrita si no fuera por muchas personas. Manuel de Pinedo, Mar Muriana y Jesús Palomo son los principales responsables del hecho de que pudiera decir –en español y en inglés– lo que quería decir. Mis amigos Mar y Jesús me asesoraron lingüísticamente a lo largo de los cuatro años de investigación y corrigieron gran parte de esta tesis. A Manuel, mi director, también le debo las horas que habrá pasado leyéndose las innumerables versiones de cada capítulo. Pero, más que estas horas –que seguramente fueron muchas– le agradezco que haya compartido conmigo su entusiasmo por la filosofía. Manuel estuvo presente en todas las etapas de mi investigación y su agudeza de pensamiento y su incansable disposición para hacer filosofía han guiado e inspirado este trabajo. He aprendido mucha filosofía con él y espero haber hecho alguna. También debo a Juan José Acero, a

María José Frápolli y a Neftalí Villanueva que mi doctorado haya sido tan rico filosóficamente. Las discusiones con mis compañeros de doctorado, Miguel Ángel Pérez, José Luis Liñán y Arancha San Ginés, también fueron de especial valor.

A lo largo de estos cuatro años hice dos estancias de investigación que supusieron un avance importante en la tesis. La primera en la Universidad de Sussex (Reino Unido), bajo la supervisión de Murali Ramachandran. Le agradezco el apoyo y el esfuerzo de integrarme en la comunidad filosófica de Sussex. En tal contexto, he tenido la oportunidad de conocer a Sarah Sawyer, quien ha sido fundamental para el desarrollo de varias cuestiones de este trabajo. Mi segunda estancia la realicé en la Universidad de California, Los Ángeles (USA), bajo la supervisión de Tyler Burge, quien también trató de integrarme en la vida filosófica de la UCLA. He tenido la oportunidad de conocer un gran filósofo y una gran persona. Tyler ha estado siempre dispuesto a contestar mis preguntas y ha discutido, algunas veces durante horas, varios de los argumentos que sostengo en esta tesis. Su rigor para hacer filosofía y su respecto hacia los alumnos fueron muy positivos para el desarrollo de mi trabajo.

Hablar de estas estancias sin hablar de mi pareja sería como contar la mitad de la historia. César me ha acompañado a todos estos sitios y ha hecho que mi adaptación –o, en varias ocasiones, la falta de ella– fuera superada con mucha más facilidad. Habiendo supuesto un gran esfuerzo por su parte, tanto profesionalmente, como psicológicamente, siempre estuvo a mi lado y me apoyó por completo. Ha sido mi puerto firme. He aprendido con él a compartir espacios, tiempo y, sobre todo, a ser más meticulosa. Pero hay que reconocer que escribir una tesis al lado de un diseñador (y músico) –con el talento que tiene él– fue mi constante motivación, especialmente las veces que me encontraba en medio de un desierto de ideas. A César también le debo el diseño de esta tesis y la maquetación. Agradezco también a Bea, que muy generosamente se ha puesto a pensar sobre cómo ilustrar la cubierta y finalmente me

ha obsequiado con la presente ilustración.

Les debo a mis antiguos profesores y amigos, Hilan Bensusan, Paulo Abrantes y Julio Cabrera, y a mis amigos y colaboradores, Makmiller Pedroso y Herivelto Souza, que haya podido hacer frente a una investigación doctoral. A los amigos que tengo en *São Paulo*, a los que hice en *Brasília* y a los ya no tan nuevos amigos de Granada, les debo la capacidad de ver desde otras perspectivas, muchas veces más perspicaces que la mía. Ellos y ellas fueron mis compañeros en este trayecto y fueron los responsables de que fuera capaz de ver más lejos que si estuviera sola, pensando desde la silla de mi despacho.

Dedico esta tesis a mis padres, Janete Borgoni y José Gonçalves, y a mi hermano, Daniel, que han tenido que superar sus *saudades* para apoyarme incondicionalmente todo el tiempo. Quisiera haberlos tenido a mi lado más que los pocos días por año en que pude estar en Brasil, pero sé que, como yo, reconocen un punto de verdad en el verso de João Pessoa de que *navegar é preciso, viver não é preciso* ("es necesario navegar, no es necesario vivir").

*

Esta tesis ha sido desarrollada en el período de disfrute de la beca pre-doctoral de formación de personal investigador (beca FPI) del actual Ministerio de Ciencia e Innovación, asociada al proyecto de investigación HUM2004-02330/FISO: Causalidad singular, contrafácticos y causación mental (Investigador Principal: Manuel de Pinedo). En los dos últimos años de investigación he tenido el apoyo de dos proyectos más, HUM2007-63797/FISO: Conocimiento, Racionalidad Acción Causal (Investigador Principal: María José García Encinas) y HUM432: Conocimiento, Verdad y Valores (Investigador Principal: Juan Antonio Nicolás). Esta investigación ha sido posible por la financiación de tales proyectos y por el apoyo del Departamento de Filosofía I de la Universidad de Granada.

PREFACE

Language, style and references

This dissertation is composed by six chapters, three written in Spanish and the other three in English. In order to confer linguistic unity to the text, I have added an appendix to each of the chapters written in Spanish: two summaries of chapters one and two, and a translation of chapter five. I have found it appropriate to locate such appendixes subsequently to their respective Spanish versions in order to maintain the line of the text's discussion. The Introduction, the Conclusion and the general Bibliographical References are only written in English.

I give bibliographical references in the form of the triad (Author, Year: Page); e.g., (Burge, 2006: 154). I refer to the year of the original publication, but the page number refers to the publication offered in the Bibliographical References, which does not always coincide with the original one. When the work is a translation, the year also refers to the publication of the original work, but the page number refers to the translation specified in the Bibliographical References, unless otherwise indicated. The only exceptions to this system are Wittgenstein's references; I use 'PI' to refer to *Philosophical Investigations* and 'Z' to *Zettel,* followed by the respective aphorism,

e.g. (PI §258). Each chapter has its own bibliographical references –in Spanish or in English, depending on the chapter's language– but the general Bibliographical References –found at the end of the dissertation– congregates all the references used in the dissertation, including the ones from the Introduction.

Three of the chapters were published or are in the process of publication, namely chapters one, four and five. The name of the philosophical journal, the volume and page numbers (when applicable) are indicated at the begging of these chapters. I have tried to preserve the material of those papers exactly as they were written when accepted for publication, but it was necessary to make some minor formal changes, such as the system of quotation and references. I have also added a few notes to those chapters which follow a different system of numeration (i, ii, iii…) from the original version (1, 2, 3…) and, instead of being footnotes, they are located at the end of the chapter, after the chapter's bibliographical references. The layout of all chapters follows the model of philosophical papers, each one with its own abstract and keywords, both in Spanish and in English.

INTRODUCTION

Knowing the nature of one's mind: an externalist basis for self-knowledge

The six chapters that compose this work were written using the format of independent papers, although dealing with the same family of problems. In this sense, they are self-contained, which in certain moments could make it difficult to perceive the continuity between them. The function of this long introduction is to make explicit the unity behind them, by emphasizing the path by which this work has been developed and the links among the main issues involved in this dissertation.

The leading idea of this work is that a thinker is what she is due to her being part of a wider reality. That means that one's thoughts about the world and one's thoughts about oneself maintain a constitutive relation to the world itself. Because of this, in order to identify psychological tokens, we shall be ready to take traits of such a reality into account. This is the essence of externalism, which is one of the two big issues of this dissertation. The other half of the story concerns issues about self-knowledge. It is not uncommon to find a line of argumentation that considers self-knowledge to be in risk once we hold the externalist nature of the mind. In fact, an intense debate about the compatibility between externalism and self-knowledge

has moved philosophy in the last twenty years. This dissertation defends a compatibilist position and also tries to understand some of the consequences externalism imposes on an approach to self-knowledge. Knowing the nature of one's mind doesn't mean having self-knowledge in the ordinary sense, but it establishes a general framework to deal with it. And my ultimate quest is to identify how self-knowledge can be understood within this framework while maintaining its central traits. But before entering into the details of this path, some terminological questions are worth discussing.

Along the dissertation, in most cases I employ the term 'externalism' (and its variations) instead of 'anti-individualism'. Nevertheless, I use both as interchangeable notions (as well as the related terms 'internalism' and 'individualism'). Burge, who coined the term 'anti-individualism' and has vastly contributed to the development of this field, points out his reasons for avoiding 'externalism' (Burge, 2006: 154): first, his use of 'individualism' is contemporary with the early uses of 'internalism', which doesn't give historical preference to any term over the other; second, 'internalism' and 'externalism' have been employed in a different and quite established context, the epistemological one; third and primarily, 'externalism' can suggest the misleading idea of spatial location, such as that the mind is outside the head. Although I doubt the present relevance of the second reason, since 'externalism' has also been amply used in Philosophy of Mind, I'm quite sensible to the last reason.

The main problem I see with locating the mind spatially is that it suggests that the mind is not possessed by the subject anymore; an idea that is rejected by this work. In the first chapter, I present the general definition of externalism I will be working with: "mental states and contents are partly individuated by external factors to one's skin". I maintain that it is crucial to recognize that the individual is spatially and temporally located, as well as most aspects of the physical and the social world that identify one's thoughts. However,

recognizing such a thing neither implies that they are external to one's mind, nor that the mind is outside the subject. What I defend is that such external factors constitute one's mind. That is why my general definition refers to external factors to *one's skin* instead of external factors to *one's mind*. In this sense, I maintain a slight difference with Burge's position once I still maintain a place to the metaphorical image that the mind is spread over the world insofar as the world, in the broad sense, is part of the nature of the mind. However, by defending that external factors constitute one's mind without being external to it, I believe I can save the main intuition behind Burge's argumentation. I insist that the mind is still a mind possessed by a subject.

My terminological option has less to do with disagreeing with Burge in the above issue than with another sort of reasons. My option has, first, a genealogical root. In its early stages, this work was influenced by Davidson in a much stronger degree than by Burge. This will be noticed along the dissertation. When Burge's philosophy really entered the scene, I was already using Davidson's terminology. This has compelled me to preserve the term 'externalism' also by practical matters. Since this work has been developed through a continuous method, some of the papers were already written when the problem really arose. My second reason for giving preference to 'externalism' concerns the starting point of this work: the investigation of the differences among externalist positions. Burge is right in stressing that 'externalism' may suggest some misleading aspect or, at least, an aspect that is not shared by all sorts of externalism. However, nowadays, most uses of the term 'anti-individualism' endorse or make reference to Burge's position itself. One significant motivation of this work was trying to understand where the different externalist positions diverge, and it seems that 'externalism' covers a wider plurality than 'anti-individualism' does. Because of this, 'externalism' seemed to offer a better basis from which to work.

Recognizing the plurality of positions under the same label of 'externalism' is something I have carried along the entire dissertation. However, I give special attention to this issue in the first chapter, where I offer a re-reading of the externalist scene. In this chapter, I search for the distinctive aspects among externalisms and I favor one position in particular. The first step in this enterprise was to rethink the difference between Putnam and Burge by weakening their difference, understood in terms of physical and social externalism respectively. The big step Burge gives in relation to Putnam is rather the expansion of the externalist results from the realm of meaning to the realm of mind.

The Twin Earth thought experiment is developed by Putnam to show how paradoxical it is to construct a notion of meaning based on the following assumptions, which he attributes to the whole tradition of theories on meaning (Putnam, 1975: 219):

(I) Knowing the meaning of a term (in the sense of being competent in its use) is a matter of being in a psychological state.

(II) The meaning of a term (its intension) determines its extension.

According to Putnam, no notion of meaning could successfully satisfy those assumptions jointly, and Twin Earth experiment is designed to reveal that. In his formulation of this experiment, Putnam defends that Oscar and Twin Oscar[1] remain in the same psychological state, while the meaning of their words 'water' varies according to the traits of their respective worlds. However, if one maintains those two conditions together, one may accept that being in a psychological state determines a term's extension, which is supposedly false in twin

1 Putnam names the inhabitant of Earth 'Oscar1' and the inhabitant of Twin Earth 'Oscar2'. I keep such a terminology only in the first chapter, where I explain Putnam's thought experiment. Along the rest of the dissertation, I refer to the inhabitant of Earth as 'Oscar' and to the inhabitant of Twin Earth as 'Twin Oscar'.

earth contexts. Between rejecting (I) or (II), Putnam concludes that (I) is false. The alleged externalist conclusion attributed to Putnam is nothing but the denial that one's psychological state determines the meaning of one's words. In the specific case of natural kind terms, what determines the meaning of one's word is its reference (apart from its stereotypes, as I will emphasize subsequently).

Four years later, Burge offered another thought experiment favoring externalism. The salient difference between his thought experiment and Putnam's is the aspect of the world chosen to vary: an aspect of the social environment –the usage of 'arthritis'– instead of a variation in the physical environment. However, Burge introduces another deep difference in relation to Putnam. He argues against Putnam's supposition that (I) needed to be rejected. Oscar and Twin Oscar were not in the same mental state if they lived in different worlds. And the point is that we don't need to consider psychological states as being narrow in the first instance. Something also stressed by Davidson (1987). In fact, in the first chapter, I give more attention to Davidson's than to Burge's argument.

In 'Other Bodies' (1982), Burge reconstructs Putnam's experiment incorporating such a criticism. He suggests that such a thought experiment could be applied to any relatively non-theoretical natural word other than 'water', and he chooses the term 'aluminum' to develop it (a term also studied by Putnam). After all, 'water' is not a very happy term to use in twin earth contexts, since it is hard to imagine everything being exactly the same but water –that Oscar and Twin Oscar are physical replicas– when we know that most part of our bodies are constituted by water. So, in order to avoid such distractions, Burge finds it appropriate to change this minor point and talk about 'aluminum' instead of 'water'. He reconstructs the experiment as follows:

Let's suppose everything that applies to the 'water' case applies to the 'aluminum' case. We have Earth, which is much the same as the world we actually live in. And we have Twin Earth, a replica

of Earth with one exception, the metal we call 'aluminum'. The element twin people call 'aluminum' shares all macro properties with the metal we have and also call 'aluminum', but those metals have different compositions. So, one could reasonably say that in Twin Earth there is no aluminum. They have twin aluminum instead. Oscar lives in Earth and Twin Oscar lives in Twin Earth. When they say or consciously think "aluminum is a light metal", the extensions of their words 'aluminum' are different, and consequently their meanings, such as Putnam has previously pointed out. Oscar's occurrences of 'aluminum' apply to aluminum and mean *aluminum*, whereas Twin Oscar's occurrences apply to twin aluminum and mean *twin aluminum*. However, Burge argues, there are more differences.

In contrast to what Putnam has noticed, Burge defends that the differences between aluminum and twin aluminum affect Oscar and Twin Oscar thoughts. Such a difference affects oblique occurrences of 'aluminum' in 'that'-clauses, which provide the content of Oscar's and Twin Oscar's mental states (Burge, 1982: 86). If, for example, Oscar believes that aluminum can be recycled, he is thinking of aluminum as aluminum. That is, Oscar is referring in thought to aluminum. When Twin Oscar believes that aluminum can be recycled, he doesn't think of aluminum as aluminum. Actually he cannot think of aluminum as aluminum since there is no aluminum in Twin Oscar's world. Twin Oscar happens to use the same word Oscar uses, but there is no way of thinking of the metal that doesn't exist in his world. They think of (refer in thought to) different stuffs. If they have such different thoughts, consequently, they are in different psychological states, which denies Putnam's supposition that Oscar and Twin Oscar could be in the same psychological state.[2]

In this sense, Burge seems to show that there is no way of

2 I am in debt with Burge's classes and seminars at UCLA for the formulation of this argument, which he calls 'the acquisition argument'. The general approach presented in the first pages of this Introduction also owes a lot to his teaching.

running a twin earth experiment while sustaining the supposition that Oscar and Twin Oscar remain in the same psychological state. The differences in one's world will not only affect the extension of one's terms, but mainly one's thoughts. This may suggest that a useful distinction between externalisms should be the one under the disjunction between 'semantic externalism' and 'psychological externalism'. That is, an externalism that is only about meaning and another that is also about mental states. However, the point of Burge's argumentation is that it is just misleading to suppose that Oscar and Twin Oscar's thoughts are not affected by the same aspects of the world that affect the extension of their words. There is no way of saying that their words refer to specific stuffs in the world without accepting that their thoughts do the same.

Because of this, along the dissertation, I've preferred to avoid the terms 'semantic externalism' and 'psychological externalism', despite the fact that those terms are usually found in the literature. In their place, I have opted for the distinction between 'global externalism' and 'two-factor externalism' –a distinction that I develop in the first chapter– both terms concerning the mental realm. While the former represents an externalism that doesn't accept narrow states and contents –such as Burge's position– the latter is the position that *externalizes* only a part of the mind, maintaining some extent of narrowness. In this sense, in order to save Putnam's externalist intuition at the time of "The Meaning of 'Meaning'", I have identified Putnam's use of 'meaning' as meaning *broad mental contents*, an interpretation also sustained by Davidson (1987). Hence, what Oscar and Twin Oscar share could be considered as being the narrow mental contents that identify their narrow mental states.

Strictly speaking, once we consider externalism as a thesis about the mind, we should conclude that Putnam wasn't really an externalist at the time of "The Meaning of 'Meaning'". He seems indeed to favor internalism insofar as all he conceived as being psychological states were comprehended by narrow states, having

become an externalist only some years after his seminal paper (See for example, Putnam, 1988: 73 or Putnam, 2006: xxi). However, as I have said above, in order to save the philosophical tradition of attributing to him a primordial role in the defense of externalism, and to be fair to the possibility of having a dual picture of the mind, I've considered his use of the term 'meaning' as possibly understood as *broad mental contents.* That allows to interpret him as representing what I've called 'two-factor externalism'.

Burge's argument seems to be successful in demonstrating that Putnam's supposition about Oscar and Twin Oscar was wrong. They were not in the same psychological state; hence (I) doesn't need to be rejected. Nevertheless, it seems that such an argument doesn't relieve someone's impression that there is a relevant similarity between Oscar and Twin Oscar's minds: how aluminum and twin aluminum *appear to them*. And that could be the basis for an ultimate appeal to narrow contents. In the first chapter, I recognize the possibility of being externalist in the sense of two-factor externalism, but I indicate that what is appealing about narrow contents is the maintenance of the subjective realm. However, I argue, it is not mandatory to understand such a realm as being narrow. The conditions of the experiment establish that aluminum and twin aluminum share their macro properties, which suggests that Oscar and Twin Oscar have the same qualitative experiences of them. However, that Oscar and Twin Oscar are in the same qualitative state of mind doesn't mean that such states need to be narrow. After all, the sameness of their experiences is part of the initial conditions of the experiment. Therefore, it could not be concluded that they are narrow because they are the same. However, it seems that one could not conclude either that they are not narrow without further argumentation. I defend that once the subjective realm is also populated with the same notions we use to think about the world, its dependence on the wider environment is inevitable. An alternative defense of global externalism (and rejection of two-factor externalism) is reached by following McDowell's

argumentation against the notion of experience committed to what he calls 'the highest common factor'. That is, arguing that Oscar's and Twin Oscar's qualitative experiences are also different from the very beginning. That alternative is developed in more details in the second chapter.

This is part of the final image reached in the first chapter: the difference between Putnam and Burge is represented by the difference between two-factor externalism and global externalism. And I favor the latter. As I said before, my first movement in that chapter was to weaken the common reference to the social and physical externalisms as being the relevant distinction between externalisms. I emphasize that such a distinction seems to blur the one just pointed above, the distinction between global and two-factor positions. It is out of the question that Burge introduced a new perspective in 1979 by singling out social aspects on the constitution of the mind. But giving excessive attention to that fact seems to assign a secondary importance to his real advance in relation to Putnam. Moreover, the distinction between social and physical externalisms motivates a misleading interpretation of Putnam's and Burge's positions as a whole, by identifying their externalisms with their most famous experiments, as if only the isolated factor had importance to them. However, both Putnam and Burge have always taken into account both physical and social aspects. Burge indeed has insisted in several papers (1982 and 2006, for example) that the physical environment has a primary role in the constitution of the mind.

Putnam is renowned for stressing the way aspects of the physical world affect one's words (and I am assuming that in consequence, they affect one's mind). However, he has also given a place to social matters. In "The Meaning of 'Meaning'", Putnam discusses the role experts play in determining the meaning of a word. This is the idea of the social division of linguistic labor. But there is also another factor, central to his theory of meaning, which could be characterized as being social: stereotypes. Putnam has defended indeed that meaning

is the conjunction of reference plus stereotype.

Putnam's criticisms to descriptive theories of meaning have played a fundamental role in the development of his position and, consequently, in the development of his externalism. In several works (1962, 1970, 1973), he has emphasized that all available definitions of a term were not able to give its meaning. In order to defend that, he has designed several cases where his criticism could be evident. In "It Ain't Necessarily So" (1962) and "Is Semantics Possible?" (1970), for example, he proposes a situation where a basic fact about cats turns to be false. He asks us to suppose that we discover that cats are not living creatures; they are instead robots remotely controlled by clever scientists from Mars. He argues that after this discovery, we still have the term 'cat' and the things we refer to with such a word. In such circumstances, we would probably readjust our knowledge about cats: instead of being living creatures, they are robots. However, if we insist on a descriptive theory of meaning, we should sustain that we were wrong in calling those things 'cats', since the meaning of 'cat' involves the property of being living creatures. But the paradoxical point is that if we maintain such a vision on meaning we lose reference and all our history of usage of the term. We should be ready to say that "the term 'cat' doesn't have a referent once cats never existed" whereas one would like to say that "the term 'cat' refers to the robots and the sentence 'cats are robots not living things' is true". Putnam insists that, in the imaginary situation about cats being discovered to be robots, "not only will we still call them 'cats', they are cats" (Putnam, 1970: 143). Putnam uses this same line of reasoning for designing other cases, such as the ones about 'tigers' and 'lemons'. He states, for example, that if it turns out that the stripes on tigers are painted on them to deceive us, we will still call them 'tigers' and they are in fact tigers; if normal lemons turn out to be blue, they are still lemons (Putnam, 1970: 143). The conclusion is that all possible descriptions still miss the function that reference has in determining meaning.

However, Putnam recognizes and stresses a significant trait about how we convey the meaning of a word. If someone asks me, for example, about the meaning of 'lemon', I will very likely show her a lemon (Putnam, 1970: 147). However, I can convey its meaning also by giving short definitions of it. And the remarkable thing is that we give those definitions in quite similar ways. If we think of 'lemon', there are a few facts that we associate with it, for example, that a lemon is a yellow citrus fruit, size of a small fist. Those facts are what Putnam calls 'stereotypes', which also compose the meaning of a term. According to Putnam, one can convey the use of 'lemon' by conveying those facts (Putnam, 1970: 148). Stereotypes are sufficient to communicate at least an approximation to the normal use of a term in spite of the fact that they are not able to give the whole meaning, since it also involves the reference that is not fixed by the subject's knowledge about it. Something is not a lemon just because it fits its central definitions, although stereotypes have an important function in communication. In fact, Putnam defends that if our stereotype of lemon changes, with time the word 'lemon' will have changed its meaning (Putnam, 1970: 148).

To sum up, Putnam criticizes the theories of meaning which hold that the meaning of a word is given by specifying a conjunction of properties. He defends that the reference plays a fundamental role in determining meaning. But the thing is that reference is not everything either. For Putnam, meaning is the conjunction of stereotype and reference. He maintains, for example, that if our stereotype of cats changes, then in time the word 'cat' will change its meaning (Putnam, 1970: 148). In the Twin Earth experiment, Oscar and Twin Oscar probably share the stereotype of water, but the meanings of their terms were different because they have different references. Stereotypes have a part in determining meaning, but they don't exhaust it. According to Putnam, linguistic competence is a matter of knowledge plus being "in the right sort of relationship to certain distinguished situations" (Putnam, 1973: 199). That is,

reference and being appropriately related to the reference is the other crucial ingredient to determine meaning. Being appropriately related to the reference means being part of a causal chain where the referent of the term in question is present, which includes being directly related to the reference or being causally linked to other individuals who were connected to it. In that sense, not only stereotypes represent a social element in Putnam's view, but also its very conception of how the referent is connected trough social causal relations. In support of Kripke's vision, Putnam states that

> [a]nyone who uses a proper name to refer is, in a sense, a member of a collective which had 'contact' with the bearer of the name: if it is surprising that a particular member of the collective need not have had such contact, and need not even have any good idea of the bearer of the name it is only surprising because we think of language as private property. (Putnam, 1973: 203)

Putnam's externalism clearly descends from Kripke's theory of meaning and from the movement of criticism to descriptive theories. However, it should be noticed that externalism is neither a theory of reference, nor necessarily involves reference. I'm supposing, in addition, that externalism is not in principle about meaning, although it certainly has some consequences on how to conceive it. There are those who identify themselves as being externalists about reference. But the problem is that being externalist in that sense doesn't seem to say much. Unless one has a somehow eccentric view about reference, it seems trivial to state that reference depends on external factors for its identification.[3] For this reason, I've stressed that I've been considering Putnam's results about 'meaning' in "The Meaning of 'Meaning'" as a result about broad mental contents. In this respect, the externalist conclusion would be that stereotypes don't determine broad mental contents completely. The way the world is plays a crucial role on that.

3 This criticism was made by Burge in discussion about those questions.

The social aspect Putnam discusses in "The Meaning of 'Meaning'" is what he calls 'the division of linguistic labor'. That is, we are competent users of certain linguistic terms even when they are better controlled by experts on certain fields of knowledge. Or, to put the point more strongly, we are able to use terms that we don't fully understand because others possess a better understanding of them (Putnam, 1975: 227). "We can rely on a special subclass of speakers" (Putnam, 1975: 228). In the case of the term 'gold', for example, "for everyone to whom gold is important for any reason has to acquire the word 'gold', but he does not have to acquire the *method of recognizing* if something is or is not gold" (Putnam, 1975: 228). And that doesn't mean that we fail to use those terms.

A related point is explored by Burge. Reliance on others is the key matter on the case of "arthritis". The patient in the thought experiment has the concept *arthritis* and employs it although she doesn't have full understanding of the term: initially, the patient mistakenly believes that she has arthritis in her thigh. However, the point is that having an incomplete understanding of 'arthritis' doesn't mean that the patient employs a different notion when she thinks or speaks of arthritis. She has a false believe precisely because she thinks of the same notion, which in this case is not completely understood by the patient. Burge successfully argues that we don't need to have complete understanding of our terms in order to be competent in using them, such as in the case of arthritis. In fact, a minimal understanding is sufficient. This is the other ingredient Burge develops to reinforce his support to the condition (I) of Putnam's paradox –"Knowing the meaning of a term (in the sense of being competent in its use) is a matter of being in a psychological state". Once Burge argues that we don't need to comprehend psychological states as being narrow, (I) is successfully sustained also because in order to be a competent user of a language, one doesn't need to have full understanding of its terms. Putnam's notion of the division of linguistic labor involves a similar intuition; however, Burge has

explicitly developed the notion of incomplete understanding applied to any linguistic term. Burge's conception of the social is relevantly different from Putnam's and needs some careful characterization.

According to Burge, the dependence of mental states on social factors lies not in convention nor is it reduced to how experts define certain terms, such as how physicians define, for example, 'arthritis'. According to Burge, if other facts about arthritis were discovered and the definition turned out to be false, we would probably keep the word and give up the original definition, even in cases of non-natural kind terms. Inspired by Quine's remarks, Burge maintains that being a definition is a passing fact[4]. Burge's account of the social has rather to do with how social interactions intermediate the connections between a subject matter and other people; that is, how interacting with others integrates us in the adequate causal connections with the subject matter. It is through those connections that we acquire the notion in question and turn ourselves competent in its use.

A possible understanding of such social causal connections is given in terms of 'deference': when I have only a minimal understanding of a certain notion, I can defer to experts who understand it better than me, and because of this I can be said to be a competent user of a notion that I don't completely understand. However, although Burge doesn't deny the phenomenon of deference, his account about reliance on others follows a very different development. The connection one has with one's community – through which one acquires concepts and to which individuation makes reference– is not conscious or under one's own control. Deference, in contrast, seems to suppose a much stronger condition than what is necessary for one to rely on others in Burge's sense. Deference seems to require being aware of one's misunderstandings

4 Although I have not found a reference where Burge explicitly defends this idea, he has repeatedly defended this idea in conversation.

or incomplete understandings, and in addition, being able to defer to a certain group of experts. However, according to Burge, one's dependence on a community lies neither in one's awareness of the level of understanding of one's terms, nor in the ability to indicate the experts who better understand such notions. Burge states: "Our reliance on others places us under standards and norms that we may not have fully mastered. Moreover, we cannot in general tell by simple reflection whether and how we depend on others" (Burge, 2006: 173). Burge insists that the relations that count to the constitution of one's mind and to the individuation of it are all causal relations, even the social ones:

> The dependence commonly is buried in the history of one's usage and in dispositions not all of which are open to reflective recognition. The main issue has to do with what objective reality we are connected to and what standards for full understanding apply to those aspects of our usage that rely on such connection. (Burge, 2006: 173)

A different perspective on the social is found in Davidson (1973, 1974, 1991). Causal connections between individuals within a society are also fundamental in his account. However, Davidson attributes to the community an enabling function to the emergence of the mind that is not found in Burge. As Burge himself recognizes, he considers that interacting with others is psychologically necessary to learn a language, yet he doesn't endorse what other philosophers (such as Davidson and Wittgenstein) defend, that there is some conceptually necessary relation between learning or having a language and being in a community (Burge, 1989: 175). In addition, we could say that in Davidson, and in some interpretations of Wittgenstein, such conceptually necessary relation with belonging to a community extends from having a language to having a mind. Wittgenstein and Davidson's externalisms will be discussed in detail in the second chapter, and the third chapter is entirely dedicated to Davidson's account.

Such a variety of conceptions on the social gives support to an additional reason to weaken the division between social and physical externalisms. The different ways of understanding such a factor are relevantly important, which makes it superfluous to locate Burge, Davidson and Wittgenstein, for example, under the same group named 'social externalism'. The point is that it is not informative enough to say that a certain position accepts that the social affects the nature of one's mind, without being more precise. And the effort to be more precise leads to the second criterion of division between externalisms that I will develop in the first paper. Apart from the distinction between global and two-factor externalism, I suggest that there is a relevant division between levels of explanation about the external aspect of the mind.

Both Burge and Davidson defend that the individual is affected by the causal relations that one has with the environment and with one's fellows. In a sense, those causal relations are the ultimate explanation of the fact that the nature of the mind is not to be understood only by reference to the individual herself. Nevertheless, Davidson's story is broader. His insistence on the role of the community as being more fundamental than what it is for Burge, gives another sense to his externalism: the mind is not self-contained because it is primarily constituted by knowledge. According to Davidson, an individual would not think without having language, and one would not have language without being in a community. However, the very social life also depends on having knowledge of the world. Davidson's triangulation reserves both to the community and to the world the place of being the necessary conditions to the emergence of the mind. And communication is evidence that such conditions were satisfied. However, they are not sufficient to explain *interpretation*. Among the conditions of interpretation –of social life– he includes the banishment of the general doubt, both about my interlocutor being an intentional being and about the meaning of her terms. That is, in Davidson's view, interpretation is only

possible because we share a great range of beliefs and in addition, a great range of them is true. Social life and, consequently, having a mind are only possible once knowledge is already present in the story.[5] Because of that, one cannot consider one's mind in isolation from one's environment. Such an explanation contrasts, for example, with Burge's, who holds that the individuation of one's mind with reference to external factors responds to the fact that the appropriate causal relations have taken place. The same applies to Putnam.

In reference to those different levels of explanation, I've defended the distinction between the extrinsic externalism and the constitutive one. That new criterion responds to the reasons why the mind should be individuated by external factors to one's skin. What I call 'extrinsic externalism' is the position which explains the externality of the mind by making reference to the fact that mental states and mental contents were caused by those external factors. And constitutive externalism explains such an externality by sustaining that the mind is constituted by knowledge. Although I've conceived them as two different kinds of externalism, I point out that in principle they are not incompatible. Indeed, Davidson is an example of someone who carries those two levels together, where the second level (which represents constitutive externalism) seems to be

5 The same Davidsonian argumentation for the necessity of knowledge could serve to argue for the necessity of ignorance, such as Borgoni and Palomo (2006) defend. According to Davidson, communication is only possible when massive error is banned; that is, we cannot understand someone's beliefs (or even that someone is stating something meaningful) while believing that all her beliefs are false. Once someone occupies a place in the world, interacting with one's fellows, one has her own perspective, which is not completely isolated from others' perspectives (See Borgoni 2006). In a sense, only a creature without perspective could be completely mistaken. In other words, only a creature that always occupies a third-person perspective –who would always look from outside interpersonal relations– could be massively mistaken or under complete ignorance. Curiously, it seems that this is the same creature that could be completely correct, once conceiving someone completely correct would be to conceive it as occupying the position of God's eyes.

subsidiary of the first one (which represents extrinsic externalism).

Under this division, Burge and Putnam share the same group, although it should be recognized that, according to Burge –and probably for Putnam too–, causality is not the whole story. Yet it seems clear that, they don't maintain what I've called 'constitutive externalism'. Apart from Davidson (see 1991) –who actually defends a sort of hybrid position– such a position is clearly defended by Williamson (2000). And I favor it because, as my terminology suggests, the extrinsic path is still not sufficient to sustain the idea that external factors are really part of the mind. It is still consistent with the idea of internal and external components, which are connected by an extrinsic relation between each other. Once we consider the mental realm as fundamentally composed by knowledge, we have all what is needed to conceive the relation between the mind and the supposed external factors to one's skin as an intrinsic one. As I have said, these two levels of explanation are not incompatible. Despite the fact that I favor constitutive externalism as a central ingredient in an externalist position, that doesn't exclude the other level of explanation. However, it seems that they don't stand so easily together. Maybe because of this, Burge (2003a: 338-39) strongly insists that an epistemological conclusion cannot emerge from a metaphysical one, although for someone such as Williamson (2000) this relation takes a different direction. For him, the internalist conception of the mind is subsidiary of a conceptual question in epistemology regarding how to understand knowledge. And investigating the relation between those two levels of explanation is the underlying intuition of the third chapter, which is exclusively about Davidson's externalism.

In that chapter, I start by raising a conflict between the thought experiment of the Swampman, offered by Davidson in order to defend a version of global externalism, and his positions about language and interpretation. The puzzling aspect of his position is that his thought experiment that allegedly favors his externalism

designs a creature that is interpreted and seems to interpret others while it misses a mind; a situation that seems to be precluded by Davidson's theory of interpretation and by his denial of the third dogma of empiricism. Such a puzzle is understood in terms of the conflict between the two levels of explanation discussed above and developed in the first chapter. Apart from the discussion of such a puzzle, my aim in the third chapter is to locate Davidson's externalist theses among his positions on mind and on interpretation.

Underlying this chapter, there is a related question concerning the methodology of thought experiments and the defense of externalism. It seems that most of the strength of Davidson's externalism lies in his argumentation for general philosophical principles. In fact, some of his attempts to delineate thought experiments were not very successful. Needless to say that such a fact cannot by itself lay down the methodology of thought experiments. Burge himself has made most of his advances in this area following such a methodology. He indeed believes that reasoning about a thought experiment is one of the best cognitive tools for defending externalism; that is better than trying to come out with general principles. In addition, the central references of externalism are Putnam's and Burge's experiments. However, there seems to exist a neglecting aspect about focusing only on thought experiments to understand externalism, which is to lose sight of the general commitments behind the very experiments. Putnam's view on meaning and psychological states and Burge's overall range of experiments are not explicit in their most famous cases. A thought experiment cannot tell by itself what those commitments are. Apart from that, there are alternative ways of defending externalism, which give preference to the use of traditional philosophical arguments instead of thought experiments, which seem to be underestimated by the literature on externalism.

These questions were part of the motivation to develop the second paper of this dissertation, where I intend to analyze four

different lines of arguments that seem to provide each of those a different route to hold externalism. Three of them have already been mentioned: the arguments from Wittgenstein (1953), from Davidson (1973 and 1974) and from McDowell (1982). The fourth argument I analyze is the rejection of the two dogmas of empiricism, from Quine (1951). Wittgenstein, Davidson and McDowell are all examples of philosophers who support externalism. And that has been taken for granted in the first chapter. However, the precise location of their arguments that in principle would favor externalism is not very clear to identify. And that was part of the general task of the second chapter. Concerning Quine, despite the fact that it is questionable whether Quine is in fact an externalist, the argument studied does seem to favor an externalist position. They all treat different matters and have extremely different consequences, but they all share a kind of negative strategy to dismiss an internalist picture of the mind.

The above issues compose the first part of the dissertation, where I attempt to deal with the plurality of externalisms and to identify the overall position favored by this work. In the first three chapters, I intend to give the deserved place to such a plurality, mainly represented by H. Putnam, T. Burge, D. Davidson, L. Wittgenstein, J. McDowell and T. Williamson. I've tried to acknowledge their specific contribution to the establishment of externalism, marking some differences between them. In the first chapter, as mentioned above, my effort is to mark the relevant distinctions between those positions and to defend my own. In the second chapter, my general aim is to analyze particular arguments favoring externalism, which are not usually the primary reference in the literature on externalism. The third paper (written with Herivelto Souza), exclusively dedicated to Davidson, has a double task: to identify in more detail the theses that support Davidson's externalism –since they are spread over several works– and to apply the division between externalisms developed in the first chapter. My intention along the second and

third chapters was not to endorse all the positions treated in them, but to attempt to be accurate in identifying them, once the specific formulation of their externalisms is still an open question.

The second part of the dissertation connects the two big issues discussed in this work, externalism and self-knowledge. In the fourth chapter, I cover the two main contexts where compatibilism between externalism and self-knowledge has been challenged: slow-switching cases and the *reductio* argument of compatibilism. And part of my solution to the latter context is influenced by my confidence on the existence of an ambiguity hidden under the label of 'externalism'. I argue that the incompatibilism reached by *reductio* arguments is only sustained if we maintain a narrow conception of externalism; to be more precise, incompatibilism holds if we suppose Putnam's externalism as the representative position in the first instance. In general, externalism is out of risk under this argument.

I analyze the *reductio* argument of incompatibilism as developed by Boghossian in 1998, which maintains that if Oscar is a compatibilist, he is in position to know *a priori* the following: i. if I have thoughts with the concept *water*, then water exists; ii. I have thoughts with the concept *water*; iii. Water exists. According to Boghossian, Oscar is able to know (i) because it is an externalist commitment; he knows (ii) by self-knowledge; and he is able to know (iii) by deduction of (i) and (ii). However, Boghossian argues, it is unacceptable to know *a priori* that water exists.

In response to such an argument, I refer to some different strategies available to the compatibilist, such as the one developed by Sawyer (1998), who argues against the very basis of the *reductio* argument: it is not intrinsically unacceptable that someone can have empirical knowledge from introspective knowledge (Sawyer, 1998: 528). However, my position against Boghossian's argument hinges on the fact that such an argument is only successful if one supposes Putnam's externalism in the first instance, because (i) is not an intrinsic externalist commitment; (i) is not even a common

commitment found in other externalist positions. In that chapter, I argue that it is Putnam's alleged commitment to atomism what could justify someone to consider statement (i) as an externalist consequence. If one reads the Twin Earth experiment as favoring the conclusion that the external trait of Oscar's mental states about water is explained by the fact that water has directly caused such a thought, (i) seems to be available. That is, if one holds that being in contact with water is necessary and sufficient for causing me a thought about water, (i) is available. However, an externalism would hardly sustain such a view; i.e., that the external aspect of one's concepts is a consequence of their being caused by some related external objects, in isolation from other factors and concepts. Indeed, we have some concepts that fail to have an object related to them. Therefore, (i) should not be considered a general externalist commitment and Boghossian's argument should not be considered as a general challenge to compatibilism. However, one might argue that not even Putnam is committed to such a view. In such a case, we could indicate that there is still an alternative way to show how Putnam's position is what supports Boghossian's argument, by showing how Putnam's conception of meaning is behind his externalism.

In "The Meaning of 'Meaning'", Putnam's first step towards externalism is to deny the idea of methodological solipsism: "the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed" (Putnam, 1975: 220). Applying such a denial to the case of water, we could understand that the psychological states that involve the concept of *water* entail the existence of the relevant individual other than the subject to whom the state is ascribed. From Putnam's conception about meaning, which is the composition of stereotype and reference, we can understand that the relevant individual that is entailed by the denial of methodological solipsism is the reference. Hence, in the particular case of water, one's being in a psychological state that involves *water*,

entails that water exists. Water is the relevant individual, since it is the reference. To sum up, if externalism is the conjunction of the denial of methodological solipsism, such as defined by Putnam, and the conception of meaning as being the composition of stereotype and reference, then (i) is perfectly entailed by such an externalism. In this case, it is correct to say that externalism entails that if I'm in a psychological state such as that I believe that water is wet, water exists.

However, probably nobody but Putnam, who presents a dubious case of externalism, maintains that. Considering only my scope of positions, such conception of externalism is not sustained by any of the externalist philosophers discussed in this work. Burge, for example, has a very reasonable response to Boghossian's incompatibilist argument stressing that water is not required to acquire the concept *water*: "water needs not exist in an individual's environment in order for the individual to think that water is such and such" (Burge, 2003: 262). One could, for example, acquire the concept *water* by theorizing about it in an environment where there is no water. Nevertheless, given my attention to the plurality among externalisms, I believe it should be stressed that the *reductio* argument certainly obliterates a variety of positions as being encompassed under Putnam's externalism, which is far from obvious. And such a diagnosis seems to be confirmed by Boghossian's own conception of externalism when he asserts that "externalism is the view that the concept expressed by a word is individuated in part by the referent of that word" (Boghossian, 1998: 207).

The second incompatibilist challenge appears under the form of the slow-switching cases. My answer to those contexts also involves the recognition of a plurality, but this time a plurality behind the notion of self-knowledge. I argue that the incompatibilism reached by slow-switching cases is sustained only in case we maintain a specific but problematical view about self-knowledge: a Cartesian picture of it, or in terms of the last chapter of this dissertation, if one intends to maintain an old detectivism model of self-knowledge.

I consider several incompatibilist versions of those cases (designed initially by Burge in 1988), giving prevalence to the one developed by Boghossian in 1989(a). I also consider several compatibilist responses to this challenge. Boghossian's argument maintains that in a situation where an individual is unconsciously slow-switched several times between Earth and Twin Earth, acquiring concepts from both situations such as *water* and *twin water*, he would not be able to discriminate only by introspection between his thoughts involving one of those concepts. Such an individual would not know what he thinks without engaging in an investigation of his environment. I recognize that Burge's solution, which appeals to the notion of basic self-knowledge, is successful insofar as one doesn't generally need to discriminate between relevant alternatives in order to know one's thoughts. However, the small group called by Burge as 'basic self-knowledge' doesn't represent all instances of privileged self-knowledge a subject has. And that means that Boghossian's argument may work if we were to consider some other examples of self-knowledge statements. In cases where discrimination is required in order to know one's own thoughts, the individual will lack privileged self-knowledge[6]. However, the point is that even conceding such an extent of soundness to Boghossian's argument, it doesn't go any further than showing that "externalism is *consistent with* a lack of self-knowledge; it does not show that externalism *implies* a lack of self-knowledge" (Warfield, 1997: 232).

I defend that showing some cases where an individual lacks some very particular instance of privileged self-knowledge, allegedly because we are considering an externalist individuation of one's mental states, doesn't prove anything else than a compatibilism between externalism and a lack of self-knowledge of a particular

6 The term 'privileged self-knowledge', which is used in chapter four, refers to the direct and non-empirical way by which we acquire at least part of our self-knowledge.

kind. But the point is that such a conclusion is not only acceptable, but also adequate to our real situation as cognitive subjects.[7] On the one hand, that we have privileged self-knowledge seems to be plainly true. But, on the other hand, it seems that we need to recognize that some range of self-knowledge is acquired through several means other than the privileged one. In this sense, there is no intrinsic problem in needing extra information in order to know some of one's own thoughts. Nevertheless, my compatibilist argument recognizes that there is one situation where Boghossian's incompatibilism would work: if one insisted that there could not be a case of self-knowledge which was not potentially knowable to the subject in a direct and non-empirical manner. That is, the thought experiment in question –where the maintenance of the externalism is the alleged reason for the failure of one instance of privileged self-knowledge– could be used to defend incompatibilism only in case we were assuming a very specific conception of self-knowledge; a conception that is committed to the idea that whatever is part of this realm must be entirely knowable both a priori and directly. That is what I identify as the 'Cartesian picture of self-knowledge'. However, there are other available models on self-knowledge that are fairly more interesting accounts.

That leads us to the third and last part of the dissertation, where self-knowledge is the protagonist. Treating externalism and self-knowledge together has automatically pushed me into the discussion

7 Here again we could make some linkages between the argument for the necessity of knowledge and of ignorance. As a matter of fact, we do ignore and are to some extent mistaken about what we think. But maybe there is a conceptual argument supporting this fact along the lines pointed by footnote 2. Denying that we could conceive ourselves as having only third-person perspective seems to be the necessary element for rejecting the idea that one could be in a complete ignorance position, as well as in a position of complete knowledge. This reasoning could be also applied to self-knowledge. In this work, I will defend that one can have neither only third-person perspective, nor only first-person perspective on oneself.

of incompatibilist arguments. Under my argumentation, those contexts have a very narrow scope of application: incompatibilism holds for Putnam's externalism, on the one hand, and for Cartesianism, on the other hand. As I have pointed out, the externalist position I favor in this work is very different from Putnam's and from the aspect that could support the entailment that "If I have *water*-thoughts, water exists". My position is not committed to conceiving the relation between the individual and her world and her fellows in terms of atomic causal relations in order to explain the externality of the mind. It is not committed to Putnam's conception of meaning either. In fact, I've been stressing that externalism is primarily about the mental. Concerning slow-switching cases, if Cartesianism were the only available account of self-knowledge, we would probably be under pressure to abandon externalism, since I'm supposing that the fact that we have self-knowledge is beyond any doubt. Because of this, in the sixth chapter I will discuss three main models on self-knowledge, favoring an account which will be in principle safe from the available incompatibilist challenges.

The incompatibilist debate plays a very significant function for this work, but my attempt in the last section, composed by two chapters, is to go further than reasoning about the compatibility between externalism and self-knowledge. Once the incompatibility challenge has been discussed, one can go further on discussing specific problems on self-knowledge departing from an externalist basis[8]. That doesn't mean that externalism will be able to provide all the necessary ingredients to develop such an account; after all, self-knowledge is a quite independent area of study. As I have pointed out, I defend that externalism in general is compatible with self-knowledge. However, such a conclusion doesn't exhaust the question about the consequences of externalism to approaches on

8 An example of a project to understand self-knowledge (introspective knowledge) departing from an externalist account can be found in Sawyer (1999).

self-knowledge. Depending on the sort of externalism one holds, those consequences may change.

The specific position defended in this dissertation appeals to some aspects that do interfere in the conception of self-knowledge I support. And two central elements among those aspects are the 'embedding condition' and 'transparency'. In the first chapter, my defense of global externalism involves both of them. As I've already mentioned, the appealing to two-factor externalism seems to be related to the maintenance of what I've called the 'subjective realm', a realm that concerns how the subject perceives one's own mind. Two-factor externalism reasons as if the subject were detached from her broad mental contents –the contents that in this case would suffice for the externalist aspect of one's mind– while she could have complete control over her narrow mental contents. However, I argue that this is a misleading picture. One's mental contents don't need to be narrow in order to be one's own mental contents. In the first chapter, I mention a possibility for defending global externalism by completely dismissing the subjective realm. However, I sustain that a global externalism doesn't need to take such a step. I argue that if the subjective realm involves second-order thoughts, such as Oscar's beliefs about his own beliefs about water, the same external condition that individuates his first-order thoughts individuates his second ones. This is the embedding condition I mentioned above, which I employ to defend that the subjective realm is also affected by the external environment, once we don't employ different notions when we think about the world and when we think about our own thoughts about the world. Curiously, this same condition –which holds that first-order thoughts, externalistically individuated, are somehow embedded in second-order thoughts– has supported the most widely accepted compatibilist answer to slow-switching cases. Such a condition has been used to show that in fact "there is no *special* problem for the achievement of self-knowledge in the fact that my first-order thinking is subject to an externalist dependence

thesis" (Davies, 2000: 391), although this fact does not explain by itself how it is that my second-order thought amounts to knowledge.

By 'subjective realm' I understand the realm composed by higher order mental states that are acquired by the subject in a special manner; a manner that I call 'first-person perspective'. In this sense, a global externalism that dismisses subjectivity could simply insist that there is no such a thing as first-person perspective. But I indicate that such a version of global externalism is as unsuccessful explaining a real subject as two-factor externalism is. The former makes the subject foreigner to herself, while the latter makes ourselves foreigners in the world. My use of the notion of *subjective realm* intends to echo the notion of *self-knowledge acquired by first-person perspective*; a notion that I analyze in the sixth chapter. In that chapter, I defend that an account of self-knowledge cannot disregard first-person perspective. However, I also defend that it cannot disregard third-person perspective either. I argue that both ways to acquire self-knowledge are necessary to understand how, for example, the revision of one's own thoughts takes place. This provides me with reasons to reject both Cartesianism –which in a sense was already rejected by externalist reasons during my discussion of incompatibilism– and Ryleanism. The former is compatible with the absence of third-person perspective and the latter is compatible with the absence of first-person perspective. I analyze three main models of explanation on first-person perspective: the detectivist model, the constitutivist and the expressivist. Each of these involves some variations that are also discussed. I defend a version of the expressivist account that incorporates transparency, the other ingredient I mentioned as a crucial element in my externalist and compatibilist defense.

Transparency –as I use it in this work, following Evans' (1982) remarks– is the condition that points out that questions about the world are transparent to questions about one's own beliefs about the world, when taken from a first-person perspective. One of the

relevant impacts of this notion is the dismantlement of the idea that asking oneself about one's own thoughts involves a movement of looking inside. If the subject looks to anything, she certainly looks outwards instead of inwards. That is, if someone asks whether I believe it is going to rain, I'll probably look through the window searching for some sign of it, for example. I will not focus my attention on an alleged internal state. Because of this, in my defense of global externalism, I consider transparency to be a crucial ingredient in order to dissolve a picture of mind as composed by internal items to which one looks in order to discover what one thinks of. I indicate that this could be another crucial ingredient behind a two-factor externalism (and, of course, behind an overall internalist picture of the mind). Once I weaken such a characterization of the mind, it seems that I strengthen global externalism without excluding first-person perspective. That is why, in some passages, I take transparency to be evidence of the external aspect of the first-person perspective. These questions are considered in the fifth chapter, which deals with Moore's paradox, but mainly in the last chapter, where I analyze self-knowledge and first-person authority.

In the fifth chapter, I examine a very specific context of discussion related to self-knowledge: Moore's paradox; the paradox emerged from statements such as "it is raining but I don't believe so" or "it is raining but I believe it is not". And the special nature of these statements is revealed by the fact that a sentence such as "It is raining but Paula doesn't believe so" is perfectly fine when compared to the absurd character of the previous ones. This seems to indicate that someone's statements involving self-attribution of beliefs respond to some singularities that are not present in one's statements about others.

In my discussion on the paradox, I give place to the idea that one's statements about the world reveal information about the person's system of beliefs, at the same time that one's self-attributions of beliefs make explicit the transparency of first-order thoughts to

the second ones. These two ideas are derived from two traditional accounts on the paradox: Moore's (1942) and Wittgenstein's (1953) accounts. Moore maintains that one's statements about the world such as "I went to the pictures last night" implies that I believe that I went, whereas Wittgenstein suggests that saying "I believe that p" is just another way of saying "p". I argue that none of those accounts are able to solve the paradox insofar as they seek to find a proper contradiction in a Moorean sentence and fail to do so. However, I also indicate that, behind these accounts, there are relevant intuitions that I retain in my own position. I argue against a third account of the paradox offered by Shoemaker, but I also try to save its intuition: Moore's paradox is not a problem circumscribed to the speech realm; thinking of a Moorean sentence is as problematical as stating it. My solution conjoins those three intuitions with a fourth element that is not present in these traditional accounts: the fact that Moorean sentences are absurd not because they involve a contradiction, but because they are irrational.

The irrationality of a Moorean sentence is revealed by the fact that one is expected to present a sort of unity that is damaged in Moorean cases. In my argument, I appeal to the Davidsonian view on radical interpretation, maintaing that if we want to interpret someone we should suppose the rationality of our interpretee. And that means to attribute to one's words a background of truths and coherence and also that there is a unity in her mind. I also employ the Davidsonian account of irrationality to characterize the phenomenon that occurs in stating or thinking a Moorean sentence. This account understands irrationality in terms of the possibility of the partition of the mind, which allows for situations where a mental state causes another mental state without being a reason of it. It gives me an additional element to hold that in the moment of interpretation, we just cannot consider someone's mind as divided. In fact, Davidson seems to recognize that the principle of charity in interpretation is opposed to such a partition (Davidson, 1982: 184).

In brief, my solution to the paradox amounts to the following picture. We are forced to take others' words as having something to do with their position in the world and as being the manifestation of how they deal with it. Besides this, we take for granted that such a person is not divided in several parts that interact independently from each other with other people and with the world. A Moorean sentence, however, fails to respect these principles of rationality. It hurts our logical intuitions, but because it hurts our self conception as persons, which involves the idea that we somehow maintain a unity in each of us. That is not to deny that we sometimes have contradictory behaviors and that we have tensions inside ourselves. However, in a case where I found myself in a conflict between what I think about me and my explicit behavior, I'd be expected to correct my beliefs about my own thoughts. This is the question related to my defense of the necessity both of the first and third-person perspectives over oneself, which I develop in the sixth chapter.

My main aim in the last chapter –of which I've already mentioned some issues–, is to give an account of first-person authority. And characterizing self-knowledge is central to this task. I argue that an individual can acquire self-knowledge both from taking a first and a third-person perspective towards oneself. But more than that, I defend the necessity of both perspectives. This position involves arguing against Cartesianism and Ryleanism insofar as both characterize self-knowledge exclusively in terms of one or another perspective. Cartesianism fails because it doesn't account for cases of ignorance and mistakes, but also, because, if Cartesianism were to incorporate ignorance and mistakes into the picture, it would need to abandon the exclusiveness of first-person perspective.

I defend that third-person perspective has a crucial role in the revision of one's own thoughts because some of the changes on one's self-perception are only possible with the help of others, through understanding and accepting others' diagnosis over oneself. Being able to understand others' mental states ascription to oneself requires

a certain ability to assume a third-person perspective. I illustrate this situation with some examples that involve reasoning in order to convince someone of the possession of a mental state that is denied by such a person. These cases probably involve the reference to one's behavior on which others base their analysis. In some of the cases when someone is wrong about herself, she will be able to change her mind if she is capable of incorporating others' observations into her view. That is, if she is able to analyze herself in the same way as others do. She is able to understand a friend's reasons and think about her own behavior by distancing herself from her usual perspective on her. And in these cases, only by doing this, she reaches self-knowledge. Therefore, I conclude, first-person perspective alone doesn't suffice to characterize a real subject that has self-knowledge and can revise one's thoughts. But I argue that third-person perspective alone is not enough either.

I appeal once more to the condition of transparency to emphasize the asymmetry between both perspectives and to argue against the incorporation of first-person perspective into the third one. The difference between ascribing mental states to others based on their behavior and ascribing mental states to oneself is precisely the difference between having and not having a first-person perspective. And transparency reveals one particularity of first-person perspective: the commitment one has to one's own judgments. By taking a third-person perspective on oneself, one could conclude that "I don't believe it's going to rain" by realizing, for example, that she has prepared her suitcase to go to the beach. However, such a reasoning doesn't offer any constraint for this same person to conclude, at the same time, that "it is going to rain" based on her perception of the weather; it is completely cloudy, for example. However, these two sentences together form a Moorean sentence and I argue (in chapter five) for the irrationality of those. The point seems to be that first-person perspective has indeed a crucial role on the rationality of a person. And in the terms of the fifth chapter,

such a perspective seems to have a role in the maintenance of the supposed unity within one's mind.

Having defended the necessity of both third-person and of first-person perspectives in order to understand self-knowledge, I'm left with the task of characterizing first-person perspective. While the definition of third-person perspective doesn't seem to involve deep disagreements (I assume that self-ascriptions made from this perspective are those which are based on external evidence, inference, analysis, or self-interpretation), the characterization of first-person does. And this question calls for another terminological matter that concerns especially the fourth and sixth chapters. As I have emphasized, there are some alternative ways to understand first-person perspective besides Cartesianism. And that was the basis of my compatibilist argument to respond to slow-switching cases. Those alternative models are discussed in the last chapter of this work. However, in the fourth chapter, instead of using the term 'first-person perspective', I use the term 'privileged self-knowledge', because it is the term generally found in the literature on the debate around incompatibilism. There, I indicate that such a term refers to the directness and non-empirical manner by which (part of our) self-knowledge is acquired; two traits that are shared by any model on first-person perspective. Because of this, I believe that 'privileged self-knowledge' is a term that applies to the same general phenomenon as 'self-knowledge acquired by first-person perspective'. Nevertheless, I still prefer to use the latter notion insofar as 'privileged self-knowledge' seems to make reference to privileged access, a notion that also receives a variety of interpretations, but that is mainly related to the Cartesian approach. The generality of 'privileged self-knowledge' is also evident when one realizes that directness and baseless –non-empirical– manner of acquisition seem to be present in any account of first-person perspective. However, they are neither sufficient conditions to characterize it, nor exclusive to this sort of knowledge. A biologist can be perfectly trained to

identify directly some specimen of plant, for example. Or even more, certain mathematical thoughts can be also directly and *a priori* reached by a trained person. Because of this, this generality presented in the fourth chapter is recovered in the last one, where each model does offer a more specific way to characterize first-person perspective, and consequently privileged self-knowledge. Therefore, applying this terminological remark, the incompatibilist debate discussed in the fourth chapter could be characterized as dealing with the compatibility between externalism and first-person perspective.

Among the available options I discuss, I favor the expressivist model as developed by Finkelstein (2008), but incorporating transparency as a fundamental element in such an account. The expressivist model avoids the problems of detectivism (such as the inability to account for ignorance and mistakes in its old version and the inability to differentiate between first and third-person perspective in its new version) and of constituvism, which places an excessive responsibility on the subject. Detectivism, in fact, is dismissed through a variety of objections among which we could locate the very conception of externalism I maintain in the work; an externalism that dissolves the image of the mind as being composed by internal objects to which one looks in order to know what one thinks. The version of expressivism I hold seems to provide a reasonable account of avowals, which I've identified as self-ascriptions of mental states made from first-person perspective. Once one dismisses the supposition that there is a gap between our avowals and the alleged mental states behind them, such as Finkelstein emphasizes, one is able to understand that avowals express mental states with truth value. Unlike a natural expression e.g. a smile, they also have an assertoric function. However, transparency doesn't need to be in discordance with such an account. We can keep the central ingredient of Moran's account (2001) –transparency– once it reveals the aspect of responsibility involved in avowing –the responsibility of self-ascribing mental states as being one's own– in an expressivist

account. Avowals, I defend, both express one's mental states and exhibit transparency. For when I avow for example that I believe that the World Championship in athletics has ended today, I'm probably thinking about the World Championship. And I do express my belief that the World Championship has ended today. My mental states have to do with how I see and interact with the world; and expressing them either by linguistic and non-linguistic behavior shows how I think; it shows, for example, what I take to be true or false.

I have mentioned that my main objective in the last chapter is to explain first-person authority; the supposition that most of the time my self-ascriptions of mental states constitute knowledge. And it is widely supposed that by explaining first-person perspective one acquires the clues to explain first-person authority. However, I argue against this common assumption. In the sixth chapter, I offer a puzzle based on this supposition and on the fact that, once accepted that we acquire self-knowledge by both first and third-person perspectives, it is misleading to say that the authority attributed to one's self-ascription responds only to the characteristics of avowals. It seems that nothing in one's statements shows that such self-ascriptions are exclusively products of first-person perspective. And my solution is to indicate that first-person authority is a person's attribute instead of an attribute of avowals.

My proposal combines a non-epistemic approach to first-person perspective with an epistemic explanation of first-person authority. First-person perspective exhibits the same degree of security that any expression does, but it does not amount to some kind of special epistemic access to one's mind. Moreover, such an allegedly security seems to be attributed to ordinary non-avowals self-attributions as well. It's not that we have some kind of special knowledge of our own minds in terms of the epistemic qualities of first-person mode of knowledge. However, I argue that the subject has an epistemic advantage that gives legitimacy to her authority over her mind. On the one hand, I defend a hybrid model of expressivism

that incorporates transparency. On the other hand, I defend that the roots of first-person authority are in fact epistemic; authority lies in the range of possibilities the subject has to know one's thoughts.

This picture of self-knowledge, first-person perspective and first-person authority is clearly not entailed only by the traits of the externalism I defend in the first part of the dissertation. However, as I've tried to indicate, it has its basis on it. The sort of global externalism I hold involves transparency and what I've called 'the embedding condition'. The constitutive aspect of it, on the other hand, suggests a commitment to the interdependency between self-knowledge and knowledge of the world, in a sense that would not be possible in an extrinsic position. One's thoughts are constituted by external aspects not because each one of its constituents has an atomic relation to its external cause, but because it is part of a net fundamentally constituted by knowledge; which doesn't mean that ignorance and mistakes are banned from such a picture. Exactly the opposite: they are indeed as fundamental as knowledge.

As it should be obvious from this introduction, the present dissertation only deals explicitly with mental states with propositional content. The main reason for this option is that the debates I engage with are all framed in terms of such states, perhaps with the exception of Wittgenstein's and Finkelstein's work. Still, even for these authors, considerations regarding propositional attitudes are relevant for the mind as a whole. In any case, I have avoided engaging in allegedly non-propositional aspects of the mind, and I take it to be outside the scope of this work to investigate whether the conclusions reached would extend to non-linguistic creatures or to every aspect of the mental life of linguistic animals. If there are good arguments to show that, say, phenomenical states are a subclass of intentional ones, then my option would turn out to be excessively prudential. If there are not, I will still think that this work is relevant for our understanding of the propositional mind, its relationship to the world and its access to itself.

1 En casa, en el mundo: el externismo global constitutivo[1]

Cristina Borgoni. To be published in *Teorema* vol. XXIII/3 2009.

ABSTRACT
This paper suggests a re-reading of the externalist's scene directed towards the defense of a specific position: constitutive global externalism. On the one hand, such a position sustains that the meanings of someone's words, as well as her own psychology, have an external character. On the other, it maintains that mental contents and states are individuated by reference to what is beyond the subject's skin because knowledge is constitutive of the mental.
We will arrive at such a position by studying the differences between externalisms and by discussing several externalist arguments and thought experiments.

KEYWORDS: types of externalism, causality, knowledge, broad mental states, Donald Davidson, Tyler Burge.

RESUMEN
Este trabajo propone una relectura del panorama externista y, a partir de

1 Agradezco a Manuel de Pinedo su apoyo en todas las etapas de este trabajo, y a Tyler Burge y al evaluador anónimo los importantes comentarios y mejoras sugeridas sobre el texto. Agradezco también a Mar Muriana su asesoramiento lingüístico con el español. Este trabajo ha estado parcialmente financiado por el proyecto de investigación HUM2007-63797/FISO (MEC).

ella, defiende una postura específica: el externismo global constitutivo. Por un lado, dicha posición sostiene que el significado de las palabras de un sujeto, así como su propia psicología, tienen un carácter externo. Por otro lado, mantiene que los estados y contenidos mentales se individúan con referencia a factores externos a la piel del sujeto porque el conocimiento es constitutivo para la mente.
El camino hacia tal posición pasará por el estudio de las diferencias entre externismos y por la discusión de una serie de argumentos y de experimentos mentales de marcado carácter pro-externista.

PALAVRAS CLAVE: tipos de externismo, causalidad, conocimiento, estados mentales amplios, Donald Davidson, Tyler Burge.

Introducción

Entre la variedad de lemas externistas disponibles, Timothy Williamson sugiere uno muy iluminador cuando dice que "la mente y el mundo son dos variables dependientes"[2] (Williamson, 2000: 5). Tal imagen, sin embargo, así como otras más conocidas, no son más que imágenes si no las llenamos de cuerpo, si no establecemos lo que queremos decir con 'mente', con 'mundo', o si no especificamos qué tipo de relación de dependencia existe entre los dos factores.

La finalidad de este trabajo es defender un tipo específico de externismo por medio de una relectura del panorama externista. A partir del análisis de una serie de argumentos y experimentos mentales, se propondrán marcos divisorios que harán posible diferenciar entre diversos modos de dar cabida al carácter externo de lo mental. A lo largo de este recurrido, se irán delineando los distintos aspectos de la posición favorecida en el artículo.

Llamaré a dicha posición, estructurada alrededor de dos tesis, 'externismo global constitutivo'. Por un lado, ésta defiende que el

2 Todas las traducciones son mías a no ser que se indique lo contrario.

significado de las palabras de un sujeto, así como su propia psicología, tienen un carácter externo. Por otro lado, mantiene que los estados y contenidos mentales se individúan con referencia a factores externos a la piel del sujeto porque el conocimiento es constitutivo para la mente. La última sección se ocupará de tal defensa.

Para empezar, expondré una distinción, habitual en la literatura, entre el externismo físico y el externismo social. Esta distinción hace referencia al tipo de factores externos al sujeto que cuentan para la individuación de sus contenidos mentales. Defenderé que tal distinción no solamente no nos sirve para hablar de dos clases de posturas externistas, sino que camufla aspectos que sí son genuinamente importantes. No captura, por ejemplo, lo que de hecho distingue la posición de Putnam de la de Burge, una diferencia que se hace manifiesta por medio del segundo criterio propuesto, que esclarece qué parte de lo mental tiene un carácter externo. Dicho criterio se desarrollará en la tercera sección y marcará la diferencia entre un externismo de dos factores y un externismo global.

En la segunda sección se propondrá un primer criterio relevante para distinguir externismos que establecerá por qué la mente debe ser concebida en términos externistas. Tal criterio dará lugar a la diferencia entre el externismo extrínseco y el constitutivo. Bajo tal división, Putnam y Burge sí compartirían rasgos comunes.

1. Externismo físico vs. externismo social

En esta sección me concentraré en una distinción que ha recibido mucha atención en la literatura: la distinción entre externismo físico y externismo social o, en términos de Davidson (2001), externismo perceptual y externismo social.[3] Pero antes propondré

3 Otros intentos recientes de discriminar diferencias entre posiciones externistas pueden encontrarse en Rudd (1997), en McKinsey (2002), en Lafont (2005) y en Hurley (1996).

una tesis general con la cual cualquier posición externista estaría comprometida. La llamaré 'TgE':

> TgE (Tesis general del Externismo): los contenidos (y estados) mentales son individuados, en parte, por medio de factores externos a la piel del sujeto.

Es habitual que se caracterice al externismo como la tesis según la cual los contenidos mentales se individúan por factores *externos a la mente*. Sin embargo, dicha definición no es tan general como TgE porque excluye una posición –como la que se defenderá en este trabajo– que mantiene que los factores externos que sirven a la individuación de lo mental no tienen que ser externos a la mente, aunque sigan siendo externos a los límites corporales del sujeto. Si ya de salida instituimos una división entre la mente y lo que está fuera de ella, parece que la empresa de entender la conexión fundamental entre mente y mundo ya presupone que uno empieza donde acaba la otra. Así, al tiempo que TgE nos permite reconocer que el mundo físico y el mundo social son externos a la piel del sujeto, nos permite mantener que no por ello son externos a la mente, sino que la constituyen. Esta posición tiene como consecuencia que la mente se extiende por el mundo sin que por ello deje de ser la mente de un sujeto. De esta manera, TgE mantiene el aspecto espacial que se suele encontrar en las diversas posiciones externistas que se confrontarán a lo largo del trabajo, pero no vuelve imperativo que entendamos la propia mente en términos de lo que está dentro o fuera de ella (aunque tampoco cierre tal posibilidad).

Otro aspecto importante de TgE es su formulación en términos positivos en contraste con otra posibilidad, también bastante utilizada, que sería negar la tesis internista, planteada por ejemplo en términos de sobreviniencia: los contenidos mentales sobrevienen localmente al sujeto (o los estados mentales de un individuo sobrevienen a sus estados físicos). Hay quienes defienden que la referencia a la sobreviniencia sería lo más neutral para definir cualquier posición

en este debate por referirse a una relación de determinación en lugar de a una de dependencia (Sawyer, 1997: 17). Sin embargo, Burge llama la atención sobre el hecho de que es posible tanto ser internista y rechazar la sobreviniencia local como ser externista y mantenerla. De un lado podríamos tener "un dualista que mantuviera que los estados mentales son independientes de lo que está fuera tanto de lo que es interno a la mente del individuo como de lo que le es disponible por medio de la reflexión" (Burge, 2006: 153). Y, de otro lado, podríamos tener alguien que defendiera que "cualquier diferencia ambiental que determine los estados mentales tiene algún impacto en los estados del cuerpo del individuo, de tal manera que preserve la sobreviniencia local" (Burge, 2006: 153). Además, lo que según Burge es más importante, externismo e internismo

> no tratan fundamentalmente sobre sobreviniencia, son sino sobre la naturaleza de los estados mentales, sobre sus condiciones correctas de individuación. Se ocupan de las condiciones explicativas asociadas con tales naturalezas, no de meras relaciones modales. (Burge, 2006: 153)

La primera división a la que TgE daría espacio surge por medio de la evaluación de uno de sus componentes: "los factores externos a la piel del sujeto". Y las dos opciones de interpretación de tal variable serían, por un lado, el ambiente físico y, por otro, el ambiente social.

Si tenemos en cuenta que los experimentos mentales de Putnam (1975) y de Burge (1979) son los argumentos más influyentes en la discusión, podemos entender más fácilmente la razón por la que se suele encontrar en ellos esta dualidad de posiciones.

El experimento mental de la Tierra Gemela (TG) nos pide que imaginemos un planeta prácticamente idéntico a la Tierra (T), a excepción de un único elemento, el agua, que difiere respecto a su composición química –en (T) es H_2O y en (TG), XYZ– siendo todas sus macropropiedades iguales. Oscar1 es habitante de (T), y Oscar2 –idéntico a Oscar1– habita (TG). Estamos en 1750. Dado tal contexto, ¿qué determina el significado del término 'agua' al

usarse por Oscar1 y Oscar2 en los mismos enunciados?

Putnam responde: los rasgos de sus mundos. "En 1750, Oscar1 y Oscar2 entendían el término 'agua' de maneras distintas, *aunque estuviesen en el mismo estado psicológico*, y aunque a sus comunidades científicas les faltasen cerca de cincuenta años para poder descubrir tales diferencias" (Putnam, 1975: 224).

El experimento mental de la artritis, a su vez, nos pide que imaginemos a una persona con un gran número de "actitudes proposicionales comúnmente atribuidas por medio de oraciones que contienen el término 'artritis' en una ocurrencia oblicua" (Burge, 1979: 26), entre las cuales figura el pensamiento falso de que tiene artritis en sus tendones. Al relatar tal temor a su médica ésta lo corrige, explicándole que la artritis es una inflamación específica de las articulaciones. Burge nos pide que imaginemos una situación contrafáctica a la anterior, donde la aplicación del término 'artritis', según la determinan los "médicos, lexicógrafos y personas informadas", incluye también casos de inflamación en los tendones. Dadas las dos situaciones, ¿de qué depende el término 'artritis' en la primera y en la segunda situación?

Dado que toda la historia física y mental no-intencional del paciente se mantiene fija, Burge sostiene que la variación de sus contenidos mentales solamente puede ser atribuida a las diferencias en sus contextos sociales (Burge, 1979: 28).

A la vista de estos dos experimentos, y de la división planteada en esta sección, podríamos entender que Putnam da lugar a lo que he llamado 'externismo físico', que se caracteriza por TgE junto con la condición

> Cmf (Condición del mundo físico): los factores externos que tienen que ver con la individuación de contenidos y estados mentales son rasgos del mundo físico.

Por otro lado, Burge daría lugar al 'externismo social', caracterizado por TgE, junto con la siguiente condición

> Cms (Condición del mundo social): los factores externos que tienen que ver con la individuación de contenidos y estados mentales son aspectos del contexto social.

Hay un sentido en que tal lectura es trivialmente verdadera. Putnam y Burge hacen alusión a tales factores cuando elaboran sus experimentos. Mientras todo se mantiene fijo, Putnam varía un elemento del ambiente físico –la composición del agua– y Burge varía un rasgo del ambiente social –la aplicación del término 'artritis'– ocasionando, cada cual, variaciones en los contenidos mentales del sujeto de los experimentos.

Sin embargo, dividir a los externismos en las dos clases en cuestión se vuelve superficial cuando nos damos cuenta de que Cmf y Cms trabajan más bien como factores complementarios que caracterizando dos posiciones distintas. Aunque sus experimentos mentales puedan sugerirlo, ni Putnam ni Burge sostienen que el externismo sea una posición según la cual los contenidos mentales dependen exclusivamente de uno u otro factor. Al contrario, ambos sugieren en sus textos que podrían variar sus experimentos de cara a abarcar otras variables distintas a las explícitamente tratadas en sus casos más conocidos.

Davidson, por ejemplo, nos ofrece un externismo donde están presentes ambos factores (aunque insista en hablar de los dos tipos de externismo por separado). Las tesis involucradas en su idea de triangulación parecen construir un ejemplo claro de un externismo que tiene en cuenta Cmf y Cms como condiciones complementarias, pero ambas necesarias. De manera resumida, Davidson (1973 y 1974) sostiene que, en la medida en que somos intérpretes e interpretados y nos atribuimos mutuamente mentes y significados, ya están establecidas las conexiones fundamentales entre nosotros, nuestra comunidad y el mundo. No tiene sentido hablar de mente, y por tanto, de contenidos mentales, en ausencia de cualquiera de los tres vértices: el individuo, la sociedad y el mundo.

Para Davidson, un individuo no tendría pensamiento sin que tuviera lenguaje que, a su vez, solamente se hace posible desde un contacto social. Una vida social que, ella misma, no ocurre sin que tenga por base conocimiento del mundo que surge del hecho de compartir objetivamente ese mismo mundo. El establecimiento de estas condiciones de posibilidades de lo mental es subsidiario del conjunto de tesis que compone la visión de Davidson sobre la interpretación radical, en la cual el principio de caridad juega un papel fundamental. Según mantiene Davidson, no puedo interpretar a nadie, ni tampoco ser interpretado, si parto de la duda de si estoy delante de un sujeto intencional. Pero tampoco podría hacerlo si no presupusiera un amplio rango de creencias compartidas entre nosotros y que, además, gran parte de ellas fueran verdaderas. Para Davidson, conocer parece ser condición de posibilidad de lo mental.

Sostener la división entre externismos en cuestión involucra una segunda crítica, quizá todavía más seria, que sería reducir el externismo de Burge al externismo social que parece resultar de su experimento de la artritis. Sin embargo, su posición es bastante más compleja que esta y en realidad nunca ha prescindido del factor ambiente físico (ver Burge 1982). Revisando su propio trabajo donde figura dicho experimento mental, Burge afirma:

> En aquel entonces, yo consideraba al ambiente físico como más fundamental que el ambiente social en la determinación de la naturaleza de los estados mentales. Más fundamental psicológica, ontogenética y filogenéticamente. Puse mi atención primero en el ambiente social porque pensaba que su papel era menos aparente, menos fácilmente reconocible. (Burge, 2006: 153)

Se puede reconocer que Burge ha puesto una atención más fina sobre el factor social de la que puso Putnam, que básicamente lo trató bajo la noción de 'división del trabajo lingüístico'. Se puede reconocer además, que de hecho Burge y Putnam guardan diferencias cruciales respecto a cómo caracterizaron sus externismos en la medida en que Burge sobrepasa los límites de lo planeado por Putnam,

extendiendo el externismo desde la esfera de los significados hasta todo lo mental. Lo que no se puede reconocer, sin embargo, es que la diferencia crucial entre ellos, y que serviría de parámetro para situar los demás externismos, se halla en la identificación de qué factores importan para la individuación de lo mental. La diferencia crucial entre Putnam y Burge concierne al alcance de sus externismos, una diferencia que se verá en la tercera sección. Sin embargo, siguiendo el criterio rechazado en la presente sección –la división de factores individuadores de lo mental– Putnam y Burge compartirían el mismo grupo, al igual que la mayoría de las posiciones externistas disponibles.

2. Externismo extrínseco vs. externismo constitutivo

Si hay alguna diferencia digna de ser señalada entre posiciones externistas ésta no está en la distinción entre externismo físico y externismo social. Por eso, podríamos intentar buscarla en otro aspecto que TgE involucra: a qué se debe que los contenidos (y estados) mentales deban ser individuados, en parte, por medio de factores externos a la piel del sujeto. La explicación de porqué la mente tiene un carácter externo –sea porque los contenidos y estados mentales fueron causados por factores externos, sea porque la mente se constituye por conocimiento– es lo que marcará la primera división propuesta en esta relectura del escenario externista. Para ello introduciré otros dos ejemplos de argumentación pró-externista, el experimento mental del Hombre del Pantano (Davidson, 1987) y las consideraciones wittgensteinianas en las *Investigaciones Filosóficas*.

En su experimento, Davidson (1987) nos pide que nos lo imaginemos en un pantano, al lado de un árbol sobre la cual cae un rayo. En ese instante, mientras su cuerpo es reducido a cenizas, el árbol se convierte en su réplica física, lo que ocurre completamente por casualidad. El Hombre del Pantano, la réplica de Davidson, se comporta exactamente como el antiguo Davidson. Se encuentra con

los amigos de Davidson, se comporta como si los reconociera, les habla como lo haría el mismo Davidson. Entonces Davidson nos pregunta: ¿hay alguna diferencia entre el verdadero Davidson y su réplica?

La respuesta es afirmativa; hay una diferencia y se halla en las historias causales de los contenidos mentales del verdadero Davidson y de su réplica. Según Davidson, la réplica en realidad no podría reconocer a los amigos del verdadero Davidson; no podría reconocer nada, ya que no ha conocido nada inicialmente. Lo que es más, el Hombre del Pantano, aún siendo una réplica de sus características físicas y de sus comportamientos, carecería de estados intencionales, porque le faltarían las historias causales para dar sentido a sus términos[4].

El segundo ejemplo es el externismo que se le puede atribuir a Wittgenstein en sus *Investigaciones Filosóficas* (1953), por ejemplo, el que se puede inferir del así llamado argumento contra el lenguaje privado (Rudd, 1997). Aunque Wittgenstein no se ve a sí mismo como favoreciendo ninguna postura teórica, dicho argumento –normalmente encontrado entre los aforismos 244 y 271– parece motivar un externismo si se lee bajo la siguiente estructura argumentativa:

1. Es una condición de posibilidad de un lenguaje que existan criterios de corrección para él.
2. Un lenguaje privado no tiene criterios de corrección.
3. Un lenguaje privado en los términos mencionados es, por lo tanto, imposible.

Es posible concebir distintas interpretaciones de tal argumento,

4 El Hombre del Pantano no tendría estados mentales porque le faltaría una historia causal. Sin embargo, la conclusión davidsoniana en su teoría de la interpretación establece que, si algo se pone en una situación de diálogo, y por lo tanto de ser interpretado por alguien, inexorablemente habría que eliminar la duda acerca de si ese algo tiene o no una mente. Hay una tensión en las posiciones de Davidson que incluso él mismo reconoce, pero no la discutiré aquí.

como el influyente escepticismo radical que encuentra Kripke (1982), pero me interesa señalar la interpretación que apunta en mayor medida hacia un externismo. Por una parte, dicho argumento claramente rechaza una posición internista, en el sentido de que la idea de lenguaje privado puede ser elaborada en términos de la oposición a TgE: los contenidos mentales privados, o los ítems del lenguaje privado, por definición, no son individuados con referencia a factores externos al individuo. Y por otra parte, si tal argumento es visto desde la perspectiva de toda la obra, teniéndose en cuenta por ejemplo el argumento de seguimiento de reglas, parece favorecer la lectura de que la comunidad es la instancia única de comprobación para nuestras aserciones. La razón por la cual un lenguaje privado prescinde de criterios de corrección se encuentra en el hecho de que cualquier criterio de corrección ha de estar dentro de una esfera pública. Así, si algo tiene el estatuto de lenguaje, incluso si hace referencia a eventos internos o subjetivos, ya es algo público.

En este sentido, la comunidad es la instancia de donde salen los criterios de corrección y, por consiguiente, las posibilidades de individuación de lo mental. Todo esto, sin que le falte objetividad a tales criterios. El papel imprescindible de la comunidad no reside en que ella sea la "dueña" de los criterios de corrección, en el sentido de estar separada del mundo, sino en que es la única que nos capacita para lidiar lingüísticamente con él. McDowell (1984a) sugiere que al pertenecer a una comunidad aprendemos a ver las reglas. Pero esto no significa que la comunidad sea la última instancia que determina la corrección, porque si así lo fuera: "[a] uno le gustaría decir: cualquier cosa que parezca ser correcta para *nosotros* es correcta. Y esto solamente significa que aquí no podemos hablar sobre "corrección" McDowell, 1984a: 49, nota 12)[5].

5 McDowell hace uso del conocido §258 de las *Investigaciones*, cambiando la perspectiva de primera persona del singular por la del plural. Si la sociedad estuviera desconectada del mundo, tampoco podríamos tener criterios de corrección.

Teniendo en cuenta estos dos argumentos pro-externistas y los tres vistos en la sección pasada, parece que podríamos empezar a delinear dos direcciones explicativas. Por un lado, estaríamos justificados a individuar lo mental con referencia a factores externos a nuestras pieles porque nuestros contenidos y estados mentales son en alguna medida causados por tales factores. Hay algo externo a mí que es causa de mi estado mental, sean estas relaciones entre un individuo y su mundo, o entre él y su comunidad. En este grupo se encontrarían no solamente el Davidson del experimento del Hombre del Pantano, sino también Burge y Putnam, aunque cada uno manteniendo sus especificidades. Davidson, por ejemplo, construye su experimento para dialogar directamente con Putnam, argumentando que a Putnam le había faltado ampliar su relato acerca de las interacciones causales entre personas y partes del mundo de forma que incluyeran la historia de tales conexiones. Todos ellos, sin embargo, comparten el rasgo de que la explicación tendrá, al fin y al cabo, que rescatar el hecho de que mi mente se constituye de manera externa porque estuve en las circunstancias correctas para la adquisición de los estados y contenidos que tengo; he estado en relaciones causales, sea con el mundo o con la comunidad, sean presentes o pasadas.

Por otro lado, tal tipo de explicación no da cuenta de un segundo nivel explicativo en el cual parece encajar la interpretación externista de Wittgenstein así como las consecuencias externistas delineadas a partir de la noción davidsoniana de triangulación. Por un lado, la interpretación externista sobre Wittgenstein parece no tener nada que ver con una apelación a relaciones causales entre factores externos a nosotros y nuestras mentes. Por otro lado, la consecuencia de la posición de Davidson es en última instancia la defensa de que la mente no está autocontenida porque tiene como base conocimiento[6]; un estado mental que ya de por sí no podría

6 La posición de Davidson es un poco más compleja porque parece involucrar

carecer de un aspecto externo.[7]

La sugerencia es, por lo tanto, considerar ambas posiciones como instancias de dos tipos de externismo, según la articulación de TgE con una de las siguientes condiciones:

> Ce1 (Condición explicativa 1): los contenidos (y estados) mentales deben ser individuados, en parte, por factores externos a la piel del sujeto porque fueron causados por ellos.
>
> Ce2 (Condición explicativa 2): los contenidos (y estados) mentales deben ser individuados, en parte, por factores externos a la piel del sujeto porque la mente está constituida por conocimiento.

De esta manera, la conjunción de TgE con Ce1 daría lugar a lo que llamaría 'externismo extrínseco' mientras que la conjunción de TgE con Ce2 daría lugar al 'externismo constitutivo'[8].[i]

El experimento mental de la Tierra Gemela invita a que entendamos que la diferencia de significado de 'agua' entre los usos de Oscar1 y Oscar2 se remite a la diferencia de relaciones causales entre los contenidos mentales de cada uno respecto a sus mundos. En un mundo de H_2O, los pensamientos serán sobre H_2O; en un mundo de XYZ, los pensamientos serán sobre XYZ. Con el Hombre del Pantano, como ha sido señalado, Davidson concibe la referencia

---

ambos niveles explicativos acerca del carácter externo, incluso en el caso de la triangulación. Esto se discutirá más adelante.

7 Williamson (2000) indica que parte de la resistencia al externismo está en el proyecto de definir el conocimiento; según Williamson, el conocimiento no es el resultado de la articulación, por medio de la justificación, de algo interno (creencia) y algo externo (verdad), sino que las nociones mismas de creencias y justificación dependen para su inteligibilidad de la noción de conocimiento.

8 Una cuestión constitutiva es una cuestión acerca de la naturaleza de algo. En este sentido, todo externismo sería constitutivo al ser una teoría acerca de la naturaleza de lo mental. Sin embargo, la dicotomía usada aquí entre 'externismo extrínsico' y 'externismo constitutivo' añade al término otra connotación al considerar que el mundo *forma parte* de lo mental.

a relaciones causales no como algo estanco sino como algo que involucra toda la historia causal que compondría lo mental. Con el experimento de la artritis quizá en un primer momento sea más difícil percibir que las relaciones sociales deban involucrar también relaciones causales, pero así lo mantiene el propio Burge: "Cualquier dependencia que el contenido lingüístico o psicológico tenga con respecto a los demás deriva de la confianza [*reliance*] en los otros basada en determinados tipos de relaciones causales con ellos" (Burge, 2006: 176).

El externismo constitutivo parece versar sobre una materia tan distinta del extrínseco que exige que se trate por separado. En la medida en que se conciba lo mental como ya necesariamente compuesto por conocimiento parece no haber necesidad extra de justificar el carácter externo de lo mental. Como ha sido expuesto en la sección anterior, cuando se usa la teoría de la triangulación de Davidson para sostener un externismo, una de las consecuencias importantes es que la mente solamente existe en tanto que haya conocimiento presente. Con Wittgenstein, tal conclusión no puede ser sacada tan rápidamente, pero si entendemos que sus argumentos dan espacio a la idea de que el individuo, en lo que se refiere a sus contenidos mentales, no puede estar desconectado de la comunidad, ni tampoco la comunidad puede estar desconectada del mundo, podríamos permitirnos decir que la objetividad constituye a lo mental. En este sentido podríamos también atribuirle a Wittgenstein la idea de que en la esfera de lo mental deba estar presente estados mentales con la característica de no poder estar desconectados del mundo.

Parece posible que ambos niveles explicativos caminen juntos, como es el caso con Davidson. Si, por un lado, podemos atribuirle la tesis de que no hay mente si no hay conocimiento, deberíamos reconocer que tal tesis es subsidiaria de su idea de que la objetividad de lo mental surge de las conexiones causales entre un individuo, su comunidad y el mundo que comparten. La cuestión, sin embargo,

es que las dos líneas explicativas parecen ser independientes una de la otra y, además, cada una parece ser suficiente para sostener un externismo. Pero antes nos ocuparemos de la segunda división relevante para nuestra discusión.

3. Externismo global vs. externismo de dos factores

La segunda división que propongo se da con respecto a qué parte de la mente tiene un carácter externo. Argumentaré que hay dos tipos de posturas externistas: las que entienden que toda la mente es externa y las que se conforman con que sólo una parte lo sea.

Tales posturas se caracterizarán por la adición a TgE de una de las siguientes condiciones:

> Cndv (Condición de la no-división): "los contenidos y estados mentales" tal y como figura en TgE hacen referencia a toda la mente. Es decir, no hay distinción entre contenidos amplios y estrechos.
>
> Cdv (Condición de la división): "los contenidos y estados mentales" tal y como figura en TgE hacen referencia a los llamados estados y contenidos amplios. Los contenidos estrechos siguen siendo internos.

De este modo, un 'externismo global' sería aquél que se compromete con TgE más Cndv y el 'externismo de dos factores' sería aquél que se compromete con TgE más Cdv.

Veinte años después de "The Meaning of 'Meaning'", Putnam reconoce no haber llevado su posición lo bastante lejos, al dejar un espacio abierto para las nociones de estados y contenidos mentales estrechos (Putnam, 1996: xxi), una crítica que ya le hizo Burge. Aunque ahora Putnam esté de acuerdo con las observaciones de Burge, su experimento en aquel momento necesitaba que una parte de la mente –los contenidos estrechos– mantuviese un carácter interno.

Recordemos la estrategia de Putnam en su experimento: partiendo de una situación en la que dos individuos no tuviesen ninguna posibilidad de acceso al "verdadero" significado de sus términos, estaríamos, sin embargo, dispuestos a decir que 'agua' en T y 'agua' en TG tendrían significados distintos, dado que sabemos que una está compuesta por H_2O y la otra por XYZ. Tal situación, según la propone Putnam, significa que aunque no hubiera una diferencia entre los "estados psicológicos" de los habitantes de los planetas, sí habría una diferencia entre sus contenidos mentales amplios porque había una diferencia entre sus mundos. Conclusión: los estados psicológicos no determinan la extensión de los términos.

Según McDowell indica, la noción de estado psicológico tal y como es usada por Putnam en aquel momento significa estado psicológico *en el sentido estrecho*, y además tal sentido agota lo que Putnam consideraba que un agente sabía sobre sus propios estados (McDowell, 1992: 277).

Para sostener tal posición, Putnam apela a un escenario que puede resultar, en parte, bastante intuitivo. El hecho de que los agentes cognitivos del experimento no puedan percibir la distinción entre H_2O y XYZ, así como el hecho de que tanto Oscar1 como Oscar2 tendrían alguna idea acerca del agua cuando se encontrasen con los dos líquidos, lleva a la suposición de que esta cierta vivencia subjetiva de Oscar1 y Oscar2 sobre el agua tendría que ser idéntica. La conclusión es que tal vivencia subjetiva nada puede tener que ver con el propio significado de 'agua'. Esta sería una imagen de lo que estoy llamando 'externismo de dos factores'. Un externismo que mantiene una parte de la mente como interna para poder afirmar el carácter externo de la otra.

El externismo global, por otra parte, podría ser introducido con referencia al experimento mental del Hombre del Pantano. Tal argumento fue en realidad explícitamente elaborado para motivar que el externismo se expandiera de los contenidos y estados amplios hacia los estrechos. En el experimento, Davidson concluye que

la diferencia entre el Davidson verdadero y su réplica, que había surgido unas horas atrás, no es solamente una cuestión de diferencia de significados, sino que tal diferencia se extiende igualmente a sus estados psicológicos. La repercusión se ve en términos de presencia y ausencia de mentalidad, porque la falta de una historia causal privaba al Hombre del Pantano de tener estados mentales. En este sentido, la división entre tipos de estados mentales parece perder incluso su sentido, ya que dada una diferencia en el mundo, tal diferencia tiene implicaciones para toda la mente.

La crítica de Davidson al experimento de Putnam es que la diferencia entre las historias causales de Oscar1 y de Oscar2 no nos permitiría suponer que pudiesen, en algún sentido, estar en el mismo estado psicológico, porque todo el historial de tales estados sería distinto. Según la metáfora que propone el mismo Davidson (1987), dos quemaduras de piel, una causada por el sol y otra no, podrían parecerse hasta el punto de ser visualmente indistinguibles. Sin embargo, tales quemaduras seguirían siendo, la primera una quemadura de sol y la otra no. Y si nos proponemos identificarlas tenemos inevitablemente que hacer referencia a cualidades extrínsecas a ellas, o sea, a sus causas.

El experimento del Hombre del Pantano no es el único que favorece un externismo global. Se ha indicado en diversas ocasiones que Burge sugiere una posición en la cual no tiene sentido hablar de contenidos estrechos y, por lo tanto, se encuadraría dentro de tal clasificación (Davidson, 1987: 20 y Recanati, 1993: 212).

Otro ejemplo podríamos encontrarlo en Wittgenstein, pero en este caso necesitaríamos un poco más de cuidado, dado que, en general, los argumentos wittgensteinianos suelen sugerir una variedad de interpretaciones. Con respecto a esto, hay por lo menos dos interpretaciones posibles de sus posiciones.

Por un lado, el argumento en contra del lenguaje privado podría ser entendido como la afirmación de que un lenguaje referente a nuestras vivencias internas es tan público como cualquier otro. Pero

también cabría la posibilidad de sostener que las propias vivencias internas podrían seguir siendo privadas. Si seguimos este segundo sentido, la Cdv parece adaptarse bien a dicho contexto, mientras que en el primero la que tiene espacio es Cndv, una vez que se entiende que no hay nada allá de dichos vocabularios. En tal interpretación, Wittgenstein nos estaría motivando a desistir de cualquier objeto interno, en el sentido de que toda la mente pase por el criterio de corrección que la argumentación mostraba como necesario para cualquier lenguaje.[9]

Hay, sin embargo, dos formas de ser externista global que deben ser resaltadas y que surgen en los pasos posteriores a la negación de la división entre estados y contenidos amplios y estrechos.

Los estados y contenidos estrechos se definen normalmente como aquellos independientes de factores externos. Sin embargo, la intuición es que tales estados hacen referencia a lo que llamamos vivencias subjetivas, es decir, a lo que normalmente se entiende cómo perteneciente al ámbito subjetivo. Por lo tanto, se abren dos opciones tras la adherencia a Cndv:

1. Mantener que incluso tal ámbito subjetivo tiene un carácter externo, defendiendo que las vivencias internas no necesitan ser tomadas como independientes de factores externos.

2. Negar la existencia de tal ámbito.

No prestaré a esta subdivisión la misma atención que he dado a las otras distinciones, pero es importante tenerla en cuenta, en especial para pensar qué tipo de autoconocimiento podría tener espacio en cada tipo de externismo.[10] Con respecto a los tres argumentos a los que me he referido como representantes de un externismo global,

9 Como me ha indicado acertadamente un evaluador anónimo, no sería adecuado decantarse por una u otra interpretación sin un examen más detallado de lo que permite el presente trabajo.

10 El conocido debate acerca de la compatibilidad entre autoconocimiento y externismo es tratado en Borgoni (2009a), donde indico las situaciones en que serían compatibles y en las que no.

Davidson claramente se inclinaría a sostener la opción (1) mientras Burge –según lo interpreta Davidson (1987: 20)– y Wittgenstein, en la primera interpretación que he dado de su argumento, parecen estar más del lado de la opción (2)[11].[ii]

4. ¿Qué externismo?

En la primera sección se ha argumentado que marcar una diferencia en términos de qué factores externos individúan la mente no nos ayudaría mucho a percibir los matices relevantes entre externismos. Si en su momento fue importante elaborar experimentos mentales poniendo el énfasis en uno u otro factor, entender al panorama externista por medio de tal división al día de hoy significaría ignorar la tendencia existente en aceptar que tanto el mundo físico como nuestro entorno social deben contar para la constitución de lo mental. Es más, significaría dejar las cuestiones más relevantes en un segundo plano. De esta forma, he sugerido en las secciones segunda y tercera otras divisiones según las cuales podríamos comprender mejor el panorama externista y su diversidad de posiciones. Usaré este mapa para indicar la posición favorecida por el texto.

Empezando con la segunda división –la distinción entre el externismo global y el externismo de dos factores– sugiero, así como lo hace Burge, que el externismo no tiene porqué, ni debería

11 Existe una clase de argumentos externistas que se auto-denominan 'externistas fenoménicos' y que, según mi clasificación, serían ejemplos de un externismo global, en el sentido de la opción (1): también el ámbito subjetivo se entendería de manera dependiente de factores externos. Churchland (1979), cómo apunta McCulloch (2003), presenta, por lo menos, dos experimentos mentales, el de las Modalidades Transpuestas y el de la Recalibración-M, que podrían ser entendidos de tal modo. El primero concluye que una diferencia con respecto a las cualidades intrínsecas de las sensaciones no determina una diferencia de significado. El segundo añade a la conclusión anterior la idea de que el vocabulario acerca de nuestra vivencia del mundo no es independiente del vocabulario acerca del propio mundo.

parar donde lo para Putnam. Para esto es interesante entender qué hay en juego cuando un externista defiende la existencia de estados estrechos.

Recanati (1993) plantea una objeción bastante fuerte a la noción de contenidos estrechos, desarrollada de la siguiente manera:

1. Definición de 'contenido estrecho': un contenido mental es estrecho si es interno al individuo e independiente del ambiente externo.
2. Tesis externista: el contenido involucra esencialmente relaciones con objetos en el mundo externo. Por lo tanto, no hay contenido que sea independiente del ambiente externo (Recanati, 1993: 211-212).

CONCLUSIÓN: dado el externismo, la idea misma de contenido estrecho es incoherente.[iii]

Si tal objeción es correcta, la tentativa de diferenciar entre un externismo de dos factores y otro global también sería incoherente, ya que el primero no existiría. Aunque se favorecerá el global en detrimento del primero, parece ser necesario reconocer –junto a Recanati– que hay un sentido en que es posible ser externista y sostener que existen contenidos estrechos (Recanati, 1993: 213). El argumento de McGinn conlleva una intuición fuerte en esta dirección. Según él, los contenidos mentales son esencialmente falibles: no hay representación sin la posibilidad de representación errónea (McGinn, 1982: 212-3). Es decir, hay una distinción fundamental entre dos aspectos independientes de las representaciones: aquello *que* es representado y *cómo* es representado. Según McGinn, para sostener tal característica de los contenidos mentales es necesario considerar el último aspecto como una propiedad intrínseca de la representación y el primero como una propiedad extrínseca relacional de la representación. De esta manera, un externista de dos factores tendría que insistir en que para ser externista sería suficiente con que algunos tipos de contenidos fuesen dependientes del ambiente externo.

La propiedad intrínseca de la representación de la que habla

McGinn sería lo que normalmente identificamos como el ingrediente subjetivo de un pensamiento. Y con esto volvemos a la intuición de Putnam. Mientras que los significados de 'agua' cambian cuando es usada por Oscar1 y Oscar2, sus experiencias subjetivas sobre el agua son las mismas.

Dicha imagen parece encajar a la perfección con lo que McDowell (1982) llama la suposición del 'máximo factor común' (MFC). Según McDowell, el MFC responde a la suposición de que existe algo en común entre la experiencia real y la mera alucinación, o en nuestro caso, entre la experiencia de Oscar1 y de Oscar2 acerca del agua.

McDowell indica que el MFC conlleva en realidad una serie de presuposiciones, entre ellas una concepción de la experiencia como constituida por un intermediario entre la mente y el mundo; una noción de experiencia donde cabría hablar de una imagen mental disponible a la conciencia de uno. McDowell argumenta a favor de una noción de experiencia que prescinda de tal intermediario, lo que le permite caracterizar una experiencia real y una alucinación como situaciones completamente distintas.

McDowell, al denunciar el MFC, ofrece bases tanto para la crítica de un determinado tipo de externismo como para la construcción de uno nuevo (Macarthur, 2003: 179), que caería dentro de lo que he llamado 'externismo global'. Sin embargo, es necesario reconocer que aunque su noción disyuntiva de la experiencia implique un externismo, y además implique un externismo global, no es verdad que el externismo global implique tal noción de experiencia. Además, no agota todo el sentido del externismo mcdowelliano.

El externismo de McDowell aparece en versiones más sofisticadas que la involucrada por su concepción disyuntiva de la percepción, por ejemplo en sus artículos sobre la noción de sentido, así como en sus discusiones de las consideraciones wittgensteinianas sobre seguimiento de reglas o incluso en *Mente y Mundo* (Ver respectivamente McDowell, 1977 y 1984b, McDowell, 1984a,

McDowell, 1994). Resulta bastante difícil dar lugar a cualquier noción de subjetividad –y, especialmente, de autoconocimiento– partiendo solamente de la noción disyuntiva de la experiencia de McDowell, pero eso no es así si tenemos como perspectiva el resto de su obra. McDowell ha insistido en que el elemento de objetividad de la noción de sentido, entendida de forma externista, no habría de identificarse con la perspectiva de tercera persona, o desde un acercamiento que dispensara la subjetividad. Para él, no tiene porqué haber una incompatibilidad entre la perspectiva de un sujeto (de primera persona) y la objetividad.

La importancia del rechazo del MFC para los propósitos del texto, más que servir de apoyo a una noción de experiencia como la que defiende McDowell, es disolver una cierta idea de la mente; una concepción de la mente que tiene que aceptar un intermediario entre ella y el mundo. Dispensar de dicho elemento no nos lleva directamente a abrazar la noción disyuntiva de la experiencia como hace McDowell, y por lo tanto, no nos lleva a abrazar el externismo que saldría de allí, pero sí nos lleva a eliminar una de las razones fuertes para insistir en los contenidos estrechos.

La apelación a la idea de un intermediario entre mente y mundo como correspondiente al contenido estrecho responde a la necesidad de mantener un ámbito subjetivo, donde es fundamental hablar de cómo le parece el mundo al propio sujeto. Sin embargo, la caracterización de los estados y contenidos estrechos como independientes del mundo es posterior a la idea de un ámbito de lo mental que podría ser llamado de subjetividad. Un ámbito caracterizado fundamentalmente por la sensación de que tenemos una cierta intimidad con nosotros mismos, y no por una supuesta independencia de factores externos. Ni es obligatorio mantener un intermediario entre mente y mundo, ni tampoco entender la subjetividad en términos de un ámbito independiente del mundo.

En la sección anterior, se discutió rápidamente que sería posible seguir hablando del ámbito subjetivo incluso en un

panorama externista global. Allí se vieron dos posibilidades de ser globales con respecto al externismo: una que considera que el ámbito subjetivo también debería tener un carácter externo; la otra que niega la existencia de dicho ámbito. La segunda salida es claramente insatisfactoria para dar espacio a la intimidad. Uno se volvería extranjero a sí mismo. Sin embargo, si se pensara que la única forma de mantener tal intimidad sería apelando a contenidos estrechos –privando a la mente de los factores externos– el alto precio aparecería en la otra punta del dilema: el precio de encontrarnos totalmente extranjeros en el mundo.

Cuando pienso que Granada es una ciudad muy bonita a pesar del frío que hace en invierno y, además, me doy cuenta de que la considero así aunque desee que el invierno pase rápido, no tengo porqué separar tales pensamientos, por ejemplo, en un grupo menos objetivo que el otro. O entre oraciones sobre el mundo y oraciones sobre el sujeto, creando bases para distinguir entre contenidos con carácter externo y contenidos independientes del ambiente. Mis pensamientos de que Granada es una ciudad muy bonita y de que en invierno hace mucho frío no ofrecen ningún problema para que un externista identifique su carácter externo. Pero tampoco los deberían presentar los del segundo grupo. Los conceptos, por ejemplo, de ciudad o de invierno involucrados en ellos no son distintos, ni pierden su carácter externo cuando forman parte, por ejemplo, del contenido de mis deseos. Es más, no parece haber incremento de intimidad si los consideramos como independientes del mundo. Al revés, la aceptación de que mis pensamientos y mis gustos están plenamente constituidos por rasgos de dónde me encuentro y con quiénes no hace sino enfatizar mi propio papel como sujeto de mis pensamientos.

Los pensamientos del primer grupo enseñan claramente rasgos del mundo, pero explicitan también rasgos sobre el propio sujeto, por ejemplo, desde dónde los piensa, su perspectiva. Hace explícitos rasgos de la historia del sujeto que hicieron posible la percepción de

la belleza de la ciudad así como la percepción de que cero grados significa hacer frío. Por otra parte, a los pensamientos del segundo grupo no les faltaría la característica de que si me pregunto acerca de mi opinión sobre Granada, lo que tengo en vista no es mi propia mente, sino Granada misma (ver Evans, 1982: 225). Tampoco es un caso aislado el de sentir frío y en seguida asomarse a la ventana para ver qué tiempo hace. Qué me parece Granada no es algo que adquiero poniendo toda la atención en mí misma.[12]

La segunda sección ha tratado de otra división entre externismos, entre el extrínseco y el constitutivo, que hace referencia a las razones para identificar contenidos y estados mentales por medio de factores externos. Se sugirió allí que los dos niveles de explicación podrían aparecer juntos, como es el caso con Davidson. Cuando llama la atención sobre la importancia que tienen las conexiones causales entre un individuo, su sociedad y el mundo para el surgimiento de la mentalidad, da cabida al nivel explicativo del externismo extrínseco:

> El profesor responde a dos cosas: la situación externa y las respuestas del estudiante. El estudiante responde a dos cosas: la situación externa y las respuestas del profesor. Todas estas relaciones son causales. Así se forma el triangulo esencial, lo que posibilita la comunicación sobre objetos y eventos compartidos. Pero también es este triangulo el que determina el contenido de las palabras y de los pensamientos del estudiante cuando se vuelven lo suficientemente complejos para merecer tales términos. (Davidson, 1990: 230)

Sin embargo, sería posible encontrar en Davidson también el nivel explicativo del externismo constitutivo, aunque tendríamos que poner énfasis en otras facetas de su posición, como su conclusión de que no podríamos estar en una situación de completo equívoco. Por un lado, sus posturas con respecto a la interpretación radical sostienen que uno no podría entender todas mis creencias y a la vez

12 En Borgoni (2008) defiendo que las oraciones, en tanto que afirmadas por alguien, explicitan a la vez algo sobre el mundo y algo sobre el sujeto que las afirma, y estudio cómo esto repercute en la comprensión de la paradoja de Moore.

creer que todas ellas fuesen falsas. Por otro lado, su denuncia del tercer dogma involucra un segundo aspecto de nuestras actividades comunicativas que es la conexión intrínseca entre mundo y visiones del mundo. No puede darse sentido a la noción de un 'esquema conceptual alternativo', porque una vez que se reconoce a alguien como poseyendo un esquema conceptual, se le atribuyen a la vez significados y mentalidad. No se la puede tomar como un ser intencional e ininteligible. Davidson concluye que para entender nuestras actividades se nos exige que aceptemos el hecho de que compartimos creencias con cualquier ser lingüístico. Pero, además, que muchas de tales creencias son verdaderas. Por lo tanto, según la visión davidsoniana, el conocimiento[13] podría ser indicado como una condición necesaria para la interpretación.

Si a veces Davidson argumenta que la segunda explicación está conectada con la primera, otras veces parece que podrían ser independientes. Si nuestras actividades interpretativas son evidencia de que tenemos conocimiento, parece que esto, por sí mismo, ya explicaría el carácter externo de la mente. Una vez dado este paso, no podría concebirse lo mental como independiente del mundo. En este sentido, parece crucial enfatizar la importancia y suficiencia del nivel explicativo que compone el externismo constitutivo frente al que compone el externismo extrínseco.

Williamson (2000) ofrece una buena inspiración en esta dirección, en la que la vía constitutiva se muestra como suficiente para un externismo. Su idea es que al considerar al conocimiento como el estado mental que antecede conceptual y metafísicamente al de creencia, escapamos de la concepción internista:

> La idea de que la creencia es conceptualmente anterior al conocimiento tiene otra fuente: la concepción internista de la mente,

13 'Conocimiento', aquí, no se refiere solamente a conocimiento de mundo. Las tesis davidsonianas son más ambiciosas, ya que exigen que el conocimiento de mundo, el conocimiento de las otras mentes y el autoconocimiento sean interdependientes (Davidson, 1991).

> y el mundo externo a la mente como dos variables independientes. La creencia es simplemente una función de la variable mente. La verdad es simplemente una función de la variable mundo externo, al menos cuando la proposición dada es acerca del mundo externo. Para un internista el conocimiento es una función de dos variables, y no de una sola de ellas; que uno sepa que está lloviendo no depende solamente del propio estado mental, un estado que es el mismo para aquellos que perciben la lluvia y para aquellos que la alucinan, ni tampoco depende solamente del estado del tiempo, un estado que es el mismo para aquellos que creen en las apariencias y para aquellos que dudan de ellas. (Williamson, 2000: 5)

Williamson sugiere que, frente a la tradición filosófica dominante, podríamos considerar al conocimiento como el estado básico desde el que se explicaría los otros, y además, como el estado fundamental de la mente. Su posición nos motivaría para que pensáramos que la creencia aspira al conocimiento, y no solamente a la verdad (Williamson, 2000: 47). Con esto, la conexión intrínseca entre mente y mundo ya estaría establecida. Tal opción involucra sin duda una diversidad de cuestiones, entre las cuales las epistemológicas son el grupo más evidente. Pero si le damos un voto de confianza, un cambio conceptual y metafísico como este claramente abriría espacio al que llamé 'externismo constitutivo'. Aún más, parece ofrecer una ventaja sobre una posición que se queda en el nivel explicativo del externismo extrínseco.

Uno de los problemas con sostener el externismo solamente por la vía extrínseca es el de dejarnos todavía a merced de la idea de que sea accidental tener todos los contenidos mentales que tengo. Que sea accidental que Oscar sepa o no cosas sobre el mundo, o que sepa o no lo que piensa. La vía constitutiva no vuelve nada necesariamente cognoscible, no garantiza ningún conocimiento en particular, ni tampoco acaba con la ignorancia. Lo único que se exige es que haya algo de conocimiento para comenzar a hablar de

lo mental[14]. La metafísica rechazada es una en la que tiene sentido hablar de componentes internos y externos, conectados entre sí por medio de alguna relación extrínseca a ellos. La vía extrínseca no nos ofrece, ella misma, ninguna base útil para cambiar tal imagen. Al considerar lo mental como fundamentalmente compuesto por conocimiento, se tiene todo lo necesario para concebir una conexión intrínseca entre la mente y factores externos al individuo.

Por eso se insistió inicialmente que el externismo que se iba a favorecer aquí mantendría que los factores externos que sirven para la individuación de lo mental no tendrían que ser externos a la mente, aunque siguiesen siendo externos a los límites corporales del sujeto. La articulación del criterio de la globalidad del externismo, manteniendo la subjetividad del sujeto, con el criterio explicativo del externismo constitutivo da lugar a una posición que tiene como consecuencia que la mente se extiende por el mundo sin que por ello deje de ser la mente de un sujeto.

Referencias Bibliográficas

Borgoni, C. 2008, "Interpretando la Paradoja de Moore: la irracionalidad de una oración mooreana", *Theoria* 23/2 (62), pp. 145-61.

Borgoni, C. 2009a, "When Externalism and Privileged Self-knowledge are Compatible and When They are Not", *Episteme NS* 29 (1), en prensa.

Burge, T. 1979, "Individualism and the Mental", en P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 21-83.

Burge, T. 1982, "Other Bodies", en T. Burge 2007, *Foundations of Mind*, Oxford, Oxford University Press, pp. 82-99.

Burge, T. 2006, "Postscript to 'Individualism and the Mental'", en T. Burge

14 Al tiempo que Williamson (2000) defiende al conocimiento como el estado mental básico y fundamental de la mente, establece la ignorancia también como condición de posibilidad de lo mental. Parece haber indicios de que lo mismo ocurre con Davidson.

2007, *Foundations of Mind*, Oxford, Oxford University Press, pp. 151-81.

Churchland, P. 1979, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press.

Davidson, D. 1973, "Radical Interpretation", en D. Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 125-39.

Davidson, D. 1974, "On the Very Idea of a Conceptual Scheme", en D. Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 183-98.

Davidson, D. 1987, "Knowing One's Own Mind", en D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 15-38.

Davidson, D. 1990, "Epistemology Externalized", en D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 193-204.

Davidson, D. 1991, "Three Varieties of Knowledge", en D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 205-20.

Davidson, D. 2001, "Externalisms", en P. Kotatko, P. Pagin & G. Segal (eds.) 2001, *Interpreting Davidson*, Stanford, CSLI Publications, pp. 1-16.

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

Hurley, S. 1996, "Varieties of Externalism" en R. Menary (ed), *The Extended Mind*, Ashgate, en prensa.

Kripke, S. 1982, *Wittgenstein on Rules and Private Language*, Oxford, Basil Blackwell.

Lafont, C. 2005, "Was Heidegger an Externalist?", *Inquiry* 48 (6), pp. 507-32.

Macarthur, D. 2003, "McDowell, Scepticism, and the 'Veil of Perception'", *Australasian Journal of Philosophy* 81 (2), pp. 175-90.

McCulloch, G. 2003, *The Life of the Mind: an essay on phenomenological externalism,* London, Routledge.

McDowell, J. 1977, "On the sense and reference of a Proper Name", *Mind* 86, pp. 159-85.

McDowell, J. 1982, "Criteria, Defeasibility, and Knowledge", en J. McDowell 1998, *Meaning, Knowledge & Reality*, Cambridge,

Harvard University Press, pp. 369-94.
McDowell, J. 1984a, "Wittgenstein on Following a Rule", en A. Miller & C. Wright (eds.) 2002, *Rule Following & Meaning*, Chesham, Acumen, pp. 45-80.
McDowell, J. 1984b, "*De Re* Senses", en J. McDowell 1998, *Meaning, Knowledge & Reality*, Cambridge, Harvard University Press, pp. 214-27.
McDowell, J. 1992, "Putnam on Mind and Meaning", en J. McDowell 1998, *Meaning, Knowledge & Reality*, Cambridge, Harvard University Press, pp. 275-91.
McDowell, J. 1994, *Mind and World*, Cambridge, Harvard University Press.
McGinn, C. 1982, "The Structure of Content" en A. Woodfield (ed.) 1982, *Thought and Object*, Oxford, Clarendon Press, pp. 207-58.
McKinsey, M. 2002, "Forms of Externalism and Privileged Access", *Philosophical Perspectives* 16, pp. 199-224.
Putnam, H. 1975, "The Meaning of 'Meaning", en H. Putnam 1975, *Mind, Language and Reality,* Philosophical Papers, vol. 2, Cambridge, Cambridge University Press, pp. 215-71.
Putnam, H. 1996, "Introduction", en A. Pessin & S. Goldberg (ed) (1996), *The Twin Earth Chronicles: twenty years of reflection on Hilary Putnam's "The meaning of 'meaning'"*, New York, Sharpe, pp. xiv-xxii.
Recanati, F. 1993, *Direct Reference: From Language to Thought,* Oxford, Basil Blackwell.
Rudd, A. 1997, "Two Types of Externalism", *The Philosophical-Quarterly* 47 (189), pp. 501-07.
Sawyer, S. 1997, *Semantic Externalism and Self Knowledge: Privileged Access to the World*, Phd Thesis, King's College London.
Williamson, T. 2000, *Knowledge and its Limits*, Oxford, Oxford University Press.
Wittgenstein, L. 1953, *Philosophical Investigations* (traducido por G. E. M. Anscombe), Oxford, Basil Blackwell, 1958.

i Sería posible trazar paralelos entre mi clasificación en términos de 'externismo extrínseco' y 'externismo constitutivo' y la clasificación que sugiere Moya (1998: 248) de 'externismo causal' y 'externismo normativo'. Lo que llamo 'externismo

extrínseco' se aproxima mucho a lo que Moya llama 'externismo causal': Oscar1 y Oscar2 hablan y piensan sobre líquidos distintos porque se relacionan causalmente con agua y agua gemela respectivamente. Sin embargo, el nivel explicativo del externismo constitutivo no se aproxima tan obviamente a lo que Moya llama 'externismo normativo', que se refiere al hecho de que determinadas partes del mundo externo son usadas para definir, para dar el significado de nuestras palabras (Moya, 1998: 248). "Si, con el fin de definir 'agua', digo para alguien 'agua es *esto*', la parte señalada [*ostended*] misma pasa a hacer parte del significado de la palabra y del concepto que expresa" (Moya, 1998: 248). Lo que llamo 'externismo constitutivo' se refiere al hecho de que, dado que la mente es comprensible como tal solamente bajo la atribución de conocimiento, la individuación de los estados mentales de una persona tiene que hacer referencia a factores externos a ella.

ii Complemento a la nota 11: el externismo fenoménico al cual hago referencia en la nota 11 no trata exactamente de subrayar el carácter externista de lo que se suele llamar 'qualia'. En un cierto sentido, la propia definición de qualia (entendido como la noción que se atribuye a las ideas desarrolladas por Nagel, 1974 y Jackson, 1982 y 1986) parece responder a contenidos no-conceptuales estrechos. Sin embargo, hay otros trabajos que sí intentan ofrecer una lectura externista de los qualia, como por ejemplo Lycan (2001). Aunque no trato de dicho tema en esta disertación, me parece que es posible entender las experiencias subjetivas sin hacer mención a los qualia –y, por lo tanto, evitar el tema de los contenidos no-conceptuales– como por ejemplo indicamos en Borgoni y Pedroso (2004).

iii Ver Sawyer (2007) para una defensa detallada de la inviabilidad de la noción de contenido estrecho.

Summary of Chapter:

1 At home in the world: the constitutive global externalism

(To be published in Spanish in *Teorema* XXIII/3, 2009)

ABSTRACT
This paper suggests a re-reading of the externalist's scene directed towards the defense of a specific position: constitutive global externalism. On the one hand, such a position sustains that the meanings of someone's words, as well as her own psychology, have an external character. On the other, it maintains that mental contents and states are individuated by reference to what is beyond the subject's skin because knowledge is constitutive of the mental.
We will arrive at such a position by studying the differences between externalisms and by discussing several externalist arguments and thought experiments.

KEYWORDS: types of externalism, causality, knowledge, broad mental states, Donald Davidson, Tyler Burge.

Introduction

Structure:

- Presentation of the paper's structure and its objectives. My central goals with this work are to analyze some distinctions between

externalisms and to specify the characteristics of the externalist position I hold in the paper and, consequently, along the dissertation. This is developed in four sections.

- I make explicit the working hypothesis of this paper: what people call 'externalism' responds to a variety of positions. An investigation of externalism must take the differences among externalisms into account.

1. Physical externalism vs. Social externalism

Structure:

- Identification of the general externalist thesis that will be further specified to make room for different externalisms. I call such a thesis 'TgE' ('general Externalist Thesis'): "Mental contents and states are partly individuated through external factors to one's skin".

I define this general thesis in terms of "external factors to one's skin" instead of "external factors to one's mind" because the former is broader than the latter. The latter excludes a position such as the one defended by this paper, which defends that the external factors that individuate the mental are not external to the mind itself, despite being external to the physical limits of the subject. I also justify the election of talking in terms of individuation instead of supervenience [e.g., the subject's mental states supervene (or do not supervene) on her physical states], by Burge's reasons (2006: 153, footnote 2): it is possible to be internalist and reject local supervenience as well as to be externalist and sustain local supervenience. Moreover, according to Burge, externalism deals with the explanatory conditions associated with the nature of mental states, and not with modal relations (Burge, 2006: 153, footnote 2).

- Exposition of Twin Earth (Putnam, 1975) and Arthritis (Burge,

1979) thought experiments.

Putnam stresses how physical factors affect the meaning of one's words and Burge how social aspects affect one's mind. Given the popularity of these experiments, the attention on the type of external factors seems to motivate the distinction between physical and social externalism.

- Interpretation of the difference between physical and social externalisms in terms of TgE, depending on which kind of external factors to one's skin is in place.

Physical externalism is identified by TgE plus a condition that I call 'Cmf' ('Physical world condition'): "the external factors that have to do with the individuation of mental contents and states are traits of the physical world".

Social externalism is identified by TgE plus a condition that I call 'Cms' ('Social world condition'): "the external factors that have to do with the individuation of mental contents and states are aspects of the social world".

- Weakening the importance of dividing externalisms in terms of physical and social externalism for the following reasons:

i. If such a distinction makes reference to the election between which factor –the physical world or the society– should count for the individuation of mental contents and states, the distinction is superficial. The majority of externalist positions accept that both the world and society contribute to the constitution of the mind. In terms of the scope of this dissertation, all the externalisms studied (i.e., Putnam's, Burge's, Davidson's, McDowell's, Wittgenstein's and Williamson's positions) consider Cmf and Cms as complementary conditions. At this stage of the paper, I also expose Davidson's triangulation idea which clearly endorses that both conditions are required for the individuation of mental states and contents (see chapters 2 and 3 for a detailed exposition).

ii. If such a distinction corresponds to the relevant difference between Putnam's and Burge's externalisms, the distinction is inaccurate. Despite the fact that Twin Earth's and Arthritis' experiments emphasize one of those factors, this doesn't imply that Putnam's or Burge's externalisms are reduced to those experiments. I argue, in particular, that taking such a division to be a relevant one seems to reduce Burge's externalism to what is taken to be social externalism, which is incorrect (Burge, 1982 and 2006: 153).

iii. In relation to an alleged sort of position called 'social externalism', there are also tremendous differences regarding how to characterize the roles of the social on the constitution of one's mind (e.g., the place the social has in Burge's, Davidson's and Wittgenstein's views). In this sense, the division also fails because it is not precise enough. This third reason is only implicit in the first paper, but it has been developed in more details in the Introduction to this dissertation.

2. Extrinsic externalism vs. Constitutive externalism

Structure:

- Exposition of the Swampman thought experiment (Davidson, 1987), emphasizing the Davidsonian conclusion regarding how one's causal history affects one's entire mind (see chapters 2 and 3 for a detailed exposition).

- Exposition of Wittgenstein's considerations in the *Philosophical Investigations* (see chapter 2 for a detailed exposition), emphasizing the externalism favored by the private language argument and his considerations about following a rule.

- Proposal of a new criterion of division between externalisms that responds to the reasons for the mind to be individuated by external

factors to one's skin.

Having discussed three thought experiments (Putnam's Twin Earth, Burge's Arthritis and Davidson's Swampman) and two externalist lines of argumentation (Davidson's triangulation and Wittgenstein's arguments in *PI*), I argue that we can find two different lines of explication underlying their externalist conclusions. I suggest that there is one sense of being externalist in which the explanation of the externality of the mind makes reference to the fact that mental contents were caused by external factors. But there is another sense of externalism which explains such an externality by defending that the mind is constituted by knowledge. In this sense, I identify two explanatory conditions, which added to TgE give place to the following kinds of externalisms:

'Extrinsic Externalism', which is the conjunction of TgE and 'Ce1' ('Explanatory condition1'): "Mental contents and states must be individuated through external factors to one's skin because they were caused by such external factors".

'Constitutive Externalism', which is the conjunction of TgE and 'Ce2' ('Explanatory condition2'): "Mental contents and states must be individuated through external factors to one's skin because the mind is constituted by knowledge".

-Application of this classification to the positions discussed so far, emphasizing that those two levels of explanation are not mutually exclusive.

i. Burge and Putnam belong to the same group, the extrinsic one. The general explanation we can find underlying their positions makes reference to causal relations. One's mind is constituted in a externalist manner because one has being in the adequate causal relations for the acquisition of the contents and states that one has. And individuating mental states requires the reference to such relations.

Twin Earth's thought experiment, for example, motivates the

idea that the difference between Oscar's and Twin Oscar's usages of 'water' correspond to the difference between their causal relations to the respective stuff present in each of the worlds. In a H_2O-world, words and thoughts will be about H_2O. In a XYZ-world, words and thoughts will be about XYZ. In the thought experiment of arthritis, such a trait is not so clearly perceived. However, Burge emphasizes that even social interactions involve causal relations: "Any dependence on others for linguistic or psychological content derives from reliance on others through certain types of causal relations to them" (Burge, 2006: 176).

ii. Davidson holds a hybrid position, since he argues for both conditions (see chapter 3 for more on this issue).

The Swampman thought experiment clearly makes reference to how the causal history affects one's mind. In the thought experiment, this trait marks the difference between the real Davidson and his replica. There is also an understanding of Davidson's triangulation idea exclusively in terms of causality. However, his thesis about language, mind and interpretation makes room for what I called 'constitutive externalism'.

According to Davidson, an individual would not think without having language, and one would not have language without being in a community. However, the very social life also depends on having knowledge of the world. Community and the world are necessary for the emergence of mind in Davidson's triangulation picture. And communication is evidence that those elements are present. However, they are not sufficient to explain interpretation. Among the conditions of interpretation –of social life– it figures the banishment of general doubt, both regarding whether my interlocutor is an intentional being and regarding the meaning of her terms. That is, interpretation is, under Davidson's view, only possible because we share a great range of beliefs and, in addition, a great range of them is true. Social life and consequently, having a mind, are only possible once knowledge is already present in the story. For this reason, one

cannot consider one's mind in isolation from one's environment. However, Davidson's hybrid position seems to suggest a hierarchy between the two levels of explanation about the externality of the mind. That the mind is only conceivable as being populated by knowledge is a subsidiary thesis of his idea that the objectivity of the mental emerges from the causal connections among the individual, the community and the world. Such issue is discussed in chapter 3 of this dissertation.

iii. Timothy Williamson defends a sort of externalism that belongs to what I call 'constitutive externalism'. I also suggest that we could understand Wittgenstein as belonging to this group.

Williamson (2000) indicates that part of the resistance to the externalism lies on the project of defining 'knowledge'. According to Williamson, once we consider knowledge as both conceptual and metaphysical prior to belief, one has more tools for defending externalism. One is able, for example, to attribute states (states of knowledge) that are already mental and that from the very beginning hold an intrinsic relation to the world. Under this view, knowledge ceases to be the articulation, through justification, of an internal item (belief) and an external one (reality). The very intelligibility of the notions of belief and justification depends on the notion of knowledge.

The classification of Wittgenstein under this type of externalism is not so immediate, but there seems to be a few facts about his position that motivate such an interpretation. First, the externalist interpretation of Wittgenstein's arguments does not have to do with causal relations. Second, his arguments make room for the idea that the connection between an individual, her community and the world is what explains the objectivity of the mind. In other words, such connections justify that the mental is objective. The similarity between the role objectivity has in Wittgenstein's context and the role knowledge has in Davidson's and Williamson's context is what allows us to indicate an externalist commitment in Wittgenstein.

Within the mental realm, it is necessary that there are mental states with the characteristic of not being disconnected from the world.

3. Global externalism vs. Two-factor externalism

Structure:

- Revision of the difference between Putnam and Burge in terms of the area of the mind which their externalisms apply to (see the Introduction for a more detailed discussion of this issue).

- Proposal of the second criterion of division between externalisms, corresponding to the adhesion to one of the following conditions:

'Cndv' ('Non-division condition'): "Mental contents and states", in the TgE thesis, makes reference to the entire mind. I.e., there is no distinction between broad and narrow contents.

'Cdv' ('Division condition'): "Mental contents and states", in the TgE thesis, makes reference to broad mental states and contents. Narrow contents and states are still internal.

Global externalism is identified as TgE plus Cndv and Two-factor externalism is identified as TgE plus Cdv.

-Application of this classification to the positions discussed so far.

i. Putnam embraces two-factor externalism.

ii. Burge, Davidson and Williamson fall under global externalism. In the fourth section, I add McDowell's externalism to this group. In the Introduction to this dissertation I emphasize Burge's criticism of Putnam's position, which belongs to 'two-factor externalism' in my classification. In the first chapter, I emphasize Davidson's criticism. He stresses that the difference between the causal history of Oscar and Twin Oscar doesn't allow us to think of them, in any sense, as being in the same psychological state, because their entire histories are different.

iii. Concerning Wittgenstein, where to locate his position is more open since it depends on the preferred interpretation of his arguments.

On the one hand, the private language argument could be understood as the idea that a language about our internal experiences is as public as any other language, and there is nothing (such as the experiences themselves) beyond language. On the other hand, such an argument could also be interpreted as the idea that although such a vocabulary is public, the very internal experiences remain private. The first interpretation accommodates Cndv and the second one accommodates Cdv. I favor the first interpretation in the second chapter of this dissertation.

- Suggestion of a subdivision within Global Externalism. Considering narrow contents as referring to the subjective realm, it is possible to imagine two open possibilities for a global externalism:

1. The subjective realm has an external character. The so called 'subjective experiences' don't need to be considered as being independent of external factors.

2. There is no such a realm. There is nothing beyond what we identify as 'second-order' beliefs, neither there is anything special about what people call 'subjective realm'.

I don't label those subdivisions, but it is possible to apply the classification to the positions I've being discussing. Davidson explicitly defends the first sense of global externalist, while we should also be flexible when classifying Wittgenstein in this respect. Davidson (1987: 20) accuses Burge of defending the second sense above, although such an identification isn't at all obvious. In a footnote, I point out the existence of a variety of externalist arguments offered under the name of 'phenomenical externalism', such as the one developed by McCulloch (2003) and Churchland (1979), which under this subdivision would represent the first sense of global externalism.

4. ¿Which Externalism?

I use the matrix suggested in the previous sections to delineate the position I favor in the final section. Structure:

- Defense of global externalism:

i. Discussion of some arguments against the very idea of narrow contents, such as the one indicated by Recanati (1993: 211-12) that reasons as follows: 1. Definition of narrow contents: a mental content is narrow if it is internal to the individual and it is independent of the external environment; 2. Externalist thesis: content essentially involves relations with world objects. Therefore, there is no content independent of the external environment; 3. Conclusion: given externalism, the very idea of narrow contents is incoherent. I recognize the importance of such arguments, but I also recognize that there is a certain intuition underlying the talk about narrow contents that is related to the need of maintaining a subjective realm. Like Recanati, I accept that it is possible to be only partially externalist, although this is not the best option.

ii. Brief analysis of the notion of subjectivity involved in two-factor externalism. Putnam's thought experiment seems to suggest that although Oscar's and Twin Oscar's broad contents differ, their subjective life remains the same. This idea can be accommodated within what McDowell (1982) calls 'the highest common factor' supposition (see chapter 2 for a detailed exposition of this argument). The subjective element shared by Oscar and Twin Oscar is similar to the element supposedly shared by someone that has a real experience of, for example, an apple, and someone else who only hallucinates one (e.g., the mental image of an apple). Besides McDowell's criticism of this supposition, which doesn't imply an agreement with his disjunctivist view of experience, I argue that narrow contents are not the unique, nor the best way to make room for subjectivity.

iii. Understanding subjectivity under global externalism.

Although there are some externalist approaches that sacrifice the subjective ingredient of the mind in order to sustain its externality, I argue that it is not necessary to do that. The Global Externalism I favor doesn't deprive ourselves of our subjectivity, but argues against the idea that narrow contents could provide a better basis for our intimacy. Characterizing narrow contents as independent from the world is a posterior step to the idea that there is a realm called 'subjective'. I refer to two conditions to make room for the subjective realm:

-The 'embedding condition': the same external condition that individuates one's first-order thoughts individuates one's second-order ones. In this sense, if second-order thoughts involve the subjective realm, the same condition applies to it. The subjective realm is also affected by the external environment since I don't employ different notions when I think about the world and when I think about my own thoughts about the world. I argue against a sort of division between objective thoughts (that in general correspond to first-order thoughts) and subjective thoughts (that in general correspond to second-order thoughts) making reference to the embedding condition.

-Transparency (Evans, 1982): questions about the world are transparent to questions about one's own beliefs about the world, when taken from a first-person perspective. One of the important impacts of this notion is the dismantlement of the idea that asking oneself about one's own thoughts involves a movement of looking inside. If the subject looks towards anything, she certainly looks outwards instead of inwards. That is, if someone asks whether I believe it is going to rain, I'll probably look through the window searching for some sign of it, for example.

- Favoring constitutive externalism:

i. Arguing for the independence of one level of explanation with respect to the other.

-Understanding Davidson's hybrid position (see chapter 3 for a detailed exposition).

As I've mentioned before, Davidson's hybrid position seems to suggest a hierarchy between the two levels of explanation about the externality of the mind. However, in several circumstances, such explanations seem to be independent from each other. If our interpretational activities give evidence that we are knowers, it seems that this very fact explains by itself that we cannot conceive the mental as independent from the world.

-Exposition of Williamson's (2000) argument for defending constitutive externalism.

He argues that once we consider knowledge as the fundamental mental state, prior both conceptually and metaphysically to belief, we have all we need to be externalists, yet this state would not be a hybrid state anymore, composed by a mental plus a world element. I don't discuss the epistemic problems involved in Williamson's propose. However, his position is a good example and can offer important motivations to think about a sort of externalism that takes as its starting point the idea that conceiving the mind is only possible when it is conceived as composed by knowledge, that is, intrinsically connected to the world.

ii. Advantages of constitutive over extrinsic externalism.

What I call 'extrinsic externalism' is still consistent with the idea of internal and external components, which are connected by an extrinsic relation among each other. The extrinsic route doesn't offer any useful basis to change such a picture. However, once we consider the mental realm as fundamentally composed by knowledge, we have all that is needed to conceive the relation between the mind and the supposed external factors to one's skin as an intrinsic one. The extrinsic path gives place to the idea that it is completely accidental that I have all the mental contents I actually have; that it is accidental that Oscar knows things about the world. The constitutive path doesn't turn anything necessarily knowable,

doesn't warrant any particular instance of knowledge, neither ends with ignorance. The only thing required by this position is that in order to talk about the mental, one must consider such a realm as populated by states of knowledge from the very beginning.

Conclusion

The conjunction of the global condition (Cndv) –one that maintains subjectivity– with the explicative condition of constitutive externalism (Ce1) makes room for a sort of externalism within which it makes sense to think of the mind as spread over the world, without its ceasing to be a subject's mind.

2 Externismo sin experimentos mentales

ABSTRACT
The aim of this paper is to think about alternative ways of defending externalism that dispense with the use of thought experiments. I will study four lines of arguments: Wittgenstein's argument against private language and his considerations about following a rule; the Quinean arguments against the two dogmas of empiricism; the Davidsonian arguments against the third dogma; and, finally, the arguments presented by J. McDowell against the notion of highest common factor. I will defend that each one of those arguments, independently from each other, enables us to get to an externalist position.

KEYWORDS: externalism about the mental, private language, following a rule, two dogmas of empiricism, the third dogma of empiricism, highest common factor.

RESUMEN
El objetivo de este trabajo es pensar en caminos alternativos para llegar al externismo sobre lo mental que prescindan del uso de experimentos mentales. Se estudiarán cuatro líneas argumentativas: los argumentos contra el lenguaje privado y acerca del seguimiento de reglas de Wittgenstein, los argumentos de W. O. Quine contra los dos dogmas del empirismo, los argumentos de D. Davidson contra el tercer dogma, y por último, la

argumentación de J. McDowell contra la idea de máximo factor común. Se defenderá que cada uno de estos argumentos es capaz de motivar, de forma independiente, una posición externista.

PALAVRAS CLAVE: externismo sobre lo mental, lenguaje privado, seguimiento de reglas, dos dogmas del empirismo, tercer dogma del empirismo, máximo factor común.

Introducción

Realizar un experimento mental es razonar acerca de un escenario imaginario con el objetivo de confirmar o rechazar alguna hipótesis o teoría (Gendler, 2005: 388). En el caso del externismo sobre lo mental, los experimentos mentales que juegan un papel importante son el experimento de la Tierra Gemela, propuesto por Putnam (1975) y el experimento mental de la artritis, propuesto por Burge (1979). Ambos experimentos persiguen establecer la tesis de que los contenidos (y estados) mentales son individuados, en parte, por medio de factores externos a la piel del sujeto. En el capítulo anterior (Borgoni, 2009b), tomé a esta tesis como la tesis general externista.[1] Sin embargo, que un experimento mental sirva para corroborar tal idea no quiere decir que tal tesis sea auto-explicativa. Es decir, un experimento mental no constituye una teoría, sino que más bien postula un caso extremo donde tal teoría se pondría a prueba. Por ello, el objetivo de este trabajo es pensar sobre el externismo de lo mental por medio de caminos alternativos a las dos referencias citadas. Tales

1 También en el capítulo anterior, establezco algunos matices de dicha tesis bajo las clasificaciones de 'externismo global' vs. 'externismo de dos factores' y 'externismo extrínseco' vs. 'externismo constitutivo'. La mayoría de las posiciones aquí discutidas fueron previamente clasificadas en tal capítulo, con excepción del argumento contra los dos dogmas que, como será visto, se adecua al externismo de dos factores.

caminos alternativos conllevan además una metodología alternativa al uso de los experimentos mentales: privilegian la argumentación filosófica[2]. Con este fin se analizarán cuatro líneas argumentativas. La intención de tratar los cuatro argumentos en conjunto responde a la idea de que el externismo puede involucrar más cuestiones que la de adherirse o no al lema de que "la mente no está en la cabeza" (un lema además altamente controvertido incluso para los externistas).

La primera línea será la wittgensteiniana. Se analizará el argumento contra el lenguaje privado y el del seguimiento de reglas. Se defenderá que son argumentos que dialogan, aunque no en el sentido que ha mantenido Kripke, lo que permitirá una doble lectura externista de los mismos. El primer argumento, considerado al margen del segundo, conduce hacia el externismo por medio de la reducción al absurdo de la concepción internista de lo mental. Por 'internismo' entenderé la negación de la tesis general externista:

2 Aunque este artículo sugiera la existencia de una diferencia metodológica relevante entre experimentos mentales y argumentos filosóficos clásicos, no es mi objetivo desarrollar dicha discusión en este trabajo. Mi objetivo es estudiar vías de defensa del externismo que no se basan en experimentos mentales y que han sido menos discutidos que los conocidos experimentos de Putnam (1975) y Burge (1979). Sin embargo, me abstengo de valorar los avances metodológicos de uno y del otro método. Gran parte de la discusión metodológica acerca de los experimentos mentales se concentra en el contraste experimento mental / experimento real, teniendo como cuestión central su legitimidad como método de adquisición de conocimiento: ¿cómo se puede aprender cosas nuevas sin aparentemente disponer de nuevos datos empíricos? Sin embargo, paralelo a esta cuestión, está el contraste experimento mental / argumento. Ver Norton (1996, 2004) para una defensa de que experimentos mentales son en realidad argumentos (deductivos e inductivos), y Brendel (2004) para una versión más débil de esta posición. Ver Brown (1991a, 1991b), Bishop (1999) y Bokulich (2001) para una defensa de la posición opuesta, según la cual los experimentos se distinguen de manera significativa de los argumentos. Bishop (1999), por ejemplo, argumenta que el hecho de que los experimentos mentales acepten interpretaciones desde perspectivas incompatibles les diferencia de los argumentos. La discusión sobre experimentos mentales concierne tanto su uso en filosofía como en las ciencias naturales. Para más sobre experimentos mentales en filosofía y en ciencia, ver Kuhn (1964), Horowitz y Massey (1991) y Gendler (2000).

los estados y contenidos mentales no son individuados, ni total ni parcialmente, por factores externos a la piel del sujeto. Sin embargo, podemos encontrar en Wittgenstein una motivación positiva para el externismo si consideramos a los dos argumentos como complementarios.

La segunda tarea será mostrar cómo los dos dogmas que rechaza Quine podría también estar en la base de una concepción internista sobre la mente y estudiar a partir de tal rechazo cómo podemos defender el externismo. Según Quine, a los dos dogmas subyace una motivación filosóficamente problemática que insiste en concebir el lenguaje en términos de dos componentes, uno fáctico y otro lingüístico. Esto abriría un espacio para que, en determinados casos, uno u otro componente pudiera ser nulo.

El rechazo del tercer dogma, el tercer argumento estudiado, nos conducirá a la idea de que, con respecto a la interpretación, no es posible percibir a alguien como un ser lingüístico, con una visión del mundo, y a la vez juzgarlo como usuario de una lengua ininteligible. El carácter externista de tal propuesta deriva de la disolución de la idea de separación entre mente y mundo.

La última línea argumentativa también denunciará un factor internista en la concepción de la mente, lo que McDowell llama 'Máximo Factor Común', que postularía un intermediario entre la mente y el mundo, común a las situaciones de acierto y de error. Tal argumento contiene dos pasos: primero, se abandona la idea de que los estados mentales de un individuo estén compuestos por imágenes mentales –intermediarios entre mente y mundo– a los cuales la mente accede para realizar cualquier actividad cognitiva y, en segundo lugar, se recomendará dejar de exigir que el sujeto sepa que sabe para que le atribuyamos conocimiento.

1. El argumento contra el lenguaje privado (AcLP) y el argumento de seguimiento de reglas (ASR)

Después de que Kripke defendiera que el "argumento contra el lenguaje privado (AcLP) real se encontraba en las secciones anteriores al aforismo 243" de las *Investigaciones Filosóficas* (Kripke, 1982: 3)[3], se convirtió en imperativo, para aquellos que quisiesen entrar en la discusión, pensar sobre su asociación con el argumento de seguimiento de reglas (ASR). Mientras que la interpretación tradicional situaba el AcLP entre los aforismos 243 y 271 de las *Investigaciones*[4], Kripke defendió que éste estaba, en realidad, en los aforismos anteriores a estos, al interpretarlo como un corolario del ASR, que se encontraría alrededor de §185. En esta sección mantendré que los dos argumentos dialogan, pero no en el sentido kripkeano y, a partir de ahí, buscaré sus consecuencias externistas.

Gran parte de la discusión sobre el AcLP se establece alrededor de la situación propuesta en §258, una situación donde se imagina a alguien escribiendo en un diario la ocurrencia de una determinada sensación privada. En su diario, tal persona escribiría el signo 'S' siempre que tuviera la sensación. Pero Wittgenstein advierte sobre lo que dicho ejercicio supondría: "(…) Las palabras de este lenguaje deben referirse a lo que sólo puede ser conocido por el hablante, a sus sensaciones inmediatas, privadas. Otro no puede, por tanto, entender este lenguaje" (PI §243).

La noción de lenguaje privado que está en juego involucra una

3 Todas las traducciones son mías excepto las de Wittgenstein (1953) y Quine (1951).

4 De acuerdo con la clasificación de Hacker, los argumentos contra el lenguaje privado se encuentran entre los aforismos §243 y §315 (Hacker, 1990: 3). Sin embargo, la discusión que desarrolla Kripke parece referirse al intervalo más estrecho que va desde §243 al §271. Hacker señala además que el término 'argumento contra el lenguaje privado' no es lo bastante preciso porque en tales secciones no hay solamente un argumento, sino varios (Hacker, 1990: 15).

variedad de cuestiones cuyo análisis pormenorizado sobrepasa el alcance de este trabajo; la idea de vivencias completamente privadas (en el sentido de que nadie que no fuera su dueña podría tener acceso a ellas), la idea del desarrollo de un lenguaje que describiera dichas vivencias en tales condiciones, y por tanto, la posibilidad de que existiera un lenguaje que sólo comprendiera quien lo hubiera inventado. Wittgenstein, al argumentar en contra del lenguaje privado, argumentará en contra de las tres nociones. Pero, además, argumentará en contra de una teoría del lenguaje en especial, la que supone que es suficiente para el establecimiento de un significado una conexión ostensiva entre una palabra y una sensación o un objeto.

La situación propuesta en §258 cuestiona la noción de lenguaje privado al llevar a sus últimas consecuencias la idea de que se podría fijar el significado de 'S' con referencia a mi sensación, en términos de darle a 'S' una definición ostensiva que lo conecte a una sensación:

> (…) Una definición sirve por cierto para establecer el significado de un signo. — Bien, esto ocurre precisamente al concentrar la atención; pues, por ese medio, me imprimo la conexión del signo con la sensación. — «Me la imprimo», no obstante, sólo puede querer decir: este proceso hace que yo me acuerde en el futuro de la conexión *correcta*. Pero en nuestro caso yo no tengo criterio alguno de corrección. Se querría decir aquí: es correcto lo que en cualquier caso me parezca correcto. Y esto sólo quiere decir que aquí no puede hablarse de 'correcto'. (PI §258)

Ha habido quienes han entendido que tal argumento señala un problema escéptico acerca de la aplicación de términos de vivencias internas. Según esta interpretación, aunque la definición ostensiva sea posible, nada puede garantizar que la conexión de la sensación S de hoy con su nombre se mantenga en la próxima aparición de la misma, porque parece faltar todavía otro hecho, aquel que garantiza la corrección de la conexión hecha hoy con la conexión futura. En este momento, el riesgo de un regreso infinito de hechos

sería inevitable. Una versión de tal escepticismo sería lo que Hacker llama 'escepticismo acerca de la memoria' (Hacker, 1990: 108). Sin embargo, así como han señalado otros comentaristas (Hacker, 1990: 108 y Gert, 1986: 420, por ejemplo), este no parece ser el punto central del argumento de Wittgenstein. Como se puede ver en el caso de §258, el problema no es saber si aplico el mismo término que aplico ahora en el fututo, o poder acordarme de cómo lo he aplicado en el pasado, sino que incluso en el caso presente no podríamos estar seguros de que se ha establecido ningún significado.

Por otra parte, tenemos la famosa interpretación de Kripke acerca del AcLP como corolario del ASR, cuya discusión se centra en los aforismos alrededor del caso propuesto en §185. En tal caso se nos pide que imaginemos a un alumno a quien se le enseña a anotar series de números cardinales. "[H]acemos que él, por ejemplo, a una orden de la forma «+n» anote series de la forma 0, n, 2n, 3n, etc.; así a la orden «+1» anota la serie de los números naturales" (PI §185). El alumno ha ensayado hasta 1000.

> Hacemos ahora que el alumno continúe una serie (pongamos «+2») por encima de 1000 – y él escribe: 1000, 1004, 1008, 1012. Le decimos: «¡Mira lo que has hecho!» — Él no nos entiende. Decimos: «Debías sumar dos; ¡mira cómo has empezado la serie!» — Él responde: «¡Sí! ¿No es correcta? Pensé que debía hacerlo así.» — O supón que dijese, señalando la serie: «¡Pero si he proseguido del mismo modo!» — De nada nos serviría decir «¿Pero es que no ves…?» — y repetirle las viejas explicaciones y ejemplos (…). (PI §185)

Según la interpretación escéptica de Kripke, el conjunto de observaciones acerca del ASR nos llevaría a la conclusión de que ninguno de los hechos potencialmente relevantes para fijar el significado de un símbolo en el repertorio de un determinado hablante (por ejemplo, hechos acerca de cómo el hablante ha usado la expresión, hechos acerca de sus disposiciones para usarla, y hechos sobre su historia cualitativa mental) lograría tal función. En realidad,

pensar en qué consiste que una expresión posea un significado sería ya algo engañoso porque ningún hecho podría dar cuenta de tal cosa (Boghossian, 1989b: 508). Para Kripke, el caso imaginario descrito en §185 sería la propuesta de una paradoja escéptica en los términos que aparecen en un aforismo posterior:

> Nuestra paradoja era ésta: una regla no podía determinar ningún curso de acción porque todo curso de acción puede hacerse concordar con la regla. La respuesta era: Si todo puede hacerse concordar con la regla, entonces también puede hacerse discordar. De donde no habría ni concordancia ni desacuerdo (…). (PI §201)

Aunque Wittgenstein continúe el aforismo diciendo "Que hay ahí un malentendido se muestra ya en que en este curso de pensamientos damos interpretación tras interpretación" (PI §201), Kripke insiste en la idea del escenario escéptico. Bajo la perspectiva de Kripke, si no hay ningún hecho según el cual una atribución de significado es verdadera o falsa, no hay ningún hecho según el cual un hablante pueda querer decir una cosa en lugar de otra a través de las expresiones de su lenguaje (Miller, 2002: 1). En el caso del AcLP, Kripke defendería, por tanto, que este sería solamente un caso específico del ASR, y que nos llevaría a las mismas conclusiones, pero aplicado a términos referentes a vivencias internas. Es decir, nada fijaría el significado del término 'S', así como tampoco fijaría el significado del signo '+2' en el caso del alumno.

La salida que Kripke encuentra a la supuesta paradoja escéptica parece sugerir el comunitarismo. Si no hay hechos semánticos la diferencia entre el parecer correcto y el ser correcto sería algo que sólo la comunidad podría decidir. Por contra, McDowell (1984a) insiste en que aunque la comunidad fije, para sí misma, la diferencia entre el parecer correcto y el ser correcto, necesitaría todavía de otra instancia capaz de decir si de hecho esa asignación *es* correcta[5].

5 Hacker (1990: 19) encuentra en esta cuestión una situación donde la diferencia entre el AcLP y el ASR es clara: "Si la discusión hasta el §202 de las *Investigaciones*

McDowell, que no está de acuerdo con la interpretación kripkeana de los argumentos wittgensteinianos, nos ofrece otra posibilidad de leerlos. McDowell se ocupa de las condiciones que motivan la percepción de la paradoja escéptica. Según he señalado, Kripke no presta atención a la continuación del §201. McDowell, sin embargo, insiste en ella:

> (…) Con ello mostramos que hay una captación de una regla que *no* es una *interpretación*, sino que se manifiesta, de caso en caso de aplicación, en lo que llamamos «seguir la regla» y en lo que llamamos «contravenirla». (PI §201)

McDowell sostiene que la paradoja de Kripke ocurre sólo si seguimos considerando el significado como una interpretación. Lo necesario, por tanto, sería abandonar la idea de que comprender siempre supone ofrecer una interpretación, y por tanto, ofrecer una manera de captar la regla distinta de la que nos lleva al problema de §185. De tal modo, si lo que está en juego en el ASR no es la imposibilidad de establecer la diferencia entre acierto y error, parece que la conclusión de Kripke acerca del AcLP tampoco se sostiene. Si, como pienso, McDowell tiene razón, AcLP no es una instancia más donde se verifica la paradoja escéptica pero aplicada a otra clase de términos lingüísticos. El AcLP parece establecer una crítica específica de la idea de que entidades mentales puedan dar sentido a nuestro lenguaje. Sin embargo, creo que es cierto que ambos argumentos se dirigen en la misma dirección en la medida que ambos rechazan la idea de lenguaje que entiende al significado y al significar como una relación unívoca entre un signo y un objeto (por ejemplo, entre una imagen mental y objectos privados). Visto eso, podríamos volver al AcLP y formularlo de la siguiente manera:

---

hubiera intentado demostrar que seguir una regla, como el comercio o el trueque, es solamente concebible en un grupo social, con ello no se habría demostrado que un lenguaje público en un grupo social no es la congruencia de lenguajes 'privados' construidos sobre definiciones ostensivas privadas".

i. Es una condición de posibilidad de un lenguaje que tenga criterios de corrección para su uso;
ii. El uso de un lenguaje privado no tiene criterios de corrección;
iii. Un lenguaje privado en los términos mencionados es, por tanto, imposible. No hay algo como un lenguaje privado porque no es un lenguaje.

La primera premisa se refiere a la idea de que poseer un significado es, en esencia, cuestión de poseer condiciones de corrección. Así, los enunciados de un lenguaje tienen significados si pueden ser verdaderos o falsos.

La segunda premisa aparece de modo claro al final de §258, cuando se establece el fracaso del intento de referir a la supuesta sensación S. El intento de señalar privadamente a una determinada sensación también privada, a la cual supuestamente se intentaría llamar 'S', nos dejaría sin criterios de corrección. Parece que la única manera de que exista algún criterio en tal caso sería que la propia sensación pudiera darme el criterio, es decir, que fuera suficiente establecer una conexión por medio de una definición ostensiva entre la sensación y el nombre que se le diera. Como he señalado, esta imagen es rechazada por Wittgenstein, no solamente aquí sino en gran parte de las *Investigaciones*, como por ejemplo al exponer el ASR, o incluso en sus primeros aforismos, que ponen en cuestión la imagen agustiniana del lenguaje.

Dadas las dos premisas, la conclusión inmediata de tal argumento es que el "concepto de un lenguaje privado es uno que, como mínimo, no puede ser defendido, y como máximo, es incoherente" (Preti, 2002: 56). Tal conclusión tiene un marcado carácter externista. La idea de lenguaje privado podría elaborarse en términos de la oposición a una postura externista: los elementos de dicho lenguaje, tal como 'S', se identifican solamente con referencia a factores internos al individuo. La propia sensación S, en el imaginado ejercicio de la ostensión privada del lenguaje privado, debería ser

necesaria y suficiente para identificar el estado mental de sentir 'S', o 'creer que uno siente S'. Por esto, si se mostrara la incoherencia de dicha posibilidad se llegaría al externismo por reducción al absurdo.

Esta sería una motivación negativa para llegar al externismo y además, sería una vía independiente del ASR. Cabría, sin embargo, la posibilidad de un camino positivo, que podría consistir en la tarea de explicar cuál sería la otra opción para entender la aplicación e institución de criterios de corrección que no usen la idea de hechos semánticos. En ese segundo camino sí tendríamos que aceptar la dependencia del AcLP con relación al ASR.

Kripke sugiere que el argumento wittgensteiniano nos conduciría al comunitarismo. Esto podría entenderse como una afirmación de que la premisa (ii) es verdadera porque cualquier criterio de corrección tendría que ser establecido por una comunidad. En una interpretación como la de Kripke, uno podría encontrar algún sentido externista en la medida en que insiste en que la tarea de identificar e individuar estados y contenidos mentales pertenece a la comunidad y nunca a un hablante de forma privada. Pero, como he señalado, tal lectura no se limita a señalar el papel de la comunidad en la actividad lingüística, sino que la dota de plenos poderes para el establecimiento de significados. El comunitarismo que Kripke encuentra en Wittgenstein podría, en principio, parecer externista. Sin embargo, también podría sugerir el refrán wittgensteiniano en la versión mcdowelliana: "Se querría decir aquí: es correcto todo lo que *nos* parezca correcto. Y esto sólo quiere decir que no puede hablarse aquí de 'correcto'" (McDowell, 1984a: 49, nota 12). Es decir, no estaríamos libres de la posibilidad de que nuestro lenguaje sobre el mundo, aunque público, no tuviera nada que ver con el mundo.

Como he intentado defender, la interpretación de Kripke acerca de los argumentos wittgensteinianos no solamente no parece ser la más satisfactoria, sino que su propia solución provoca una inquietud sobre la que McDowell llama la atención. Y esta es que si antes podíamos estar aislados de la comunidad, ahora podemos estar

aislados del mundo, solo que todos a la vez, y esta no parece ser la posición de Wittgenstein. Preti percibe algo similar:

> Del hecho de que nuestros pares en la comunidad jueguen un papel constitutivo en la determinación del contenido no se sigue que el contenido no sea el "raro" proceso mental interno que Wittgenstein quiere rechazar. (…) Quizá sea verdad que lo que determina significado o contenido debería ser, en parte, constituido por la mente de los otros –pero no se seguiría de ahí que el contenido de las otras mentes de la comunidad no fuera determinado por sus procesos mentales internos. El simple hecho de ser otro no es suficiente para impedir una concepción del significado en términos de estados internos. (Preti, 2002: 60)

Hay, sin embargo, otra forma de defender que los criterios de corrección solamente pueden surgir dentro de una esfera pública sin caer en el comunitarismo, y por tanto, sin comprometerse con que la determinación del significado sea constituida con referencia a lo interno a nuestras mentes. Esto es posible cuando percibimos que la institución y la aplicación de significados no son dos actividades distintas. Si los momentos de aplicación de significados son tan importantes para Wittgenstein es porque no pueden separarse de los momentos de institución de significado. Aquí, el externismo iría por un camino más positivo que el logrado con la acusación de incoherencia a la noción de lenguaje privado porque los significados se establecerían con relación a factores externos a la piel de uno, y por así decirlo, por factores externos al conjunto de individuos que componen la comunidad.

El carácter positivo de la argumentación de Wittgenstein es, sin duda, el que trae consigo la disputa entre las diversas interpretaciones de sus argumentos. La disputa, por ejemplo, acerca de qué noción de significado defiende al final Wittgenstein después de rechazar las ideas ostensivas de definición de significado. Una interpretación posible es la que ofrece McDowell, que sugiere que uno aprende a ver las reglas en la medida en la que es parte de una comunidad;

la comunidad es la única instancia que nos capacita para lidiar lingüísticamente con el mundo (1984a). Prades (2006), aunque respaldando las líneas generales de la interpretación mcdowelliana, señala que algo más necesita ser dicho además de "la introducción del requisito de la comunidad para el significado lingüístico como una consecuencia del requisito de que debe haber una manera de captar el significado que no es una interpretación" (Prades, 2006: 150); e indica que tal elemento adicional tendría que ver con el papel que el comportamiento expresivo juega en la genealogía del contenido (Prades, 2006: 149; 151)[6]. Finkelstein (2008) subraya un elemento relacionado en la posición de Wittgenstein, pero antes de atribuir un papel central a la expresión, señala que el paso fundamental en Wittgenstein es la disolución de la idea de que existe un golfo, un hueco, entre las palabras y sus significados (o entre intenciones y acciones, o incluso entre expresiones y estados mentales). Finkelstein defiende que la gran lección que podemos aprender de la discusión sobre seguir reglas es que no hay tal golfo. Según esta interpretación, la paradoja del §201 es disuelta por Wittgenstein al rechazar el impulso de ver tal hueco; entre, por ejemplo una orden y su ejecución (una imagen sugerida en §431). De tal manera, Finkelstein logra dar cabida a la idea de que el malentendido de la paradoja del §201 se debe a que se venía entendiendo que el captar una regla era dar una interpretación. Como señala Finkelstein (2008: 81), dar una interpretación tiene sentido solamente en los casos en los que una persona se haya equivocado o en los que hay un peligro real de

6 Si por una parte, el papel de la comunidad surge del rechazo de toda la "mitología filosófica" acerca de la comprensión (como interpretación) –y en esto McDowell lleva razón– por otra parte, este paso no sería suficiente para mostrar que "el significado lingüístico no-comunitario no es posible" (Prades, 2006: 151). Según Prades (2006: 151), este paso sería suficiente si conllevara "otro corolario de la crítica de Wittgenstein a las teorías tradicionales de la intencionalidad: es porque el contenido no se fija por medio de la interpretación de rasgos no representacionales, por lo que éste puede fijarse solamente por el comportamiento expresivo".

que se equivoque. En el caso normal, nosotros captamos la regla directamente[7].

2. Argumentos contra los Dos Dogmas (AcDD) – Willard O. v. Quine

> El empirismo moderno ha sido en gran parte condicionado por dos dogmas. Uno de ellos es la creencia en cierta distinción fundamental entre verdades que son *analíticas*, basadas en significaciones, con independencia de consideraciones fácticas, y verdades que son sintéticas, basadas en los hechos. El otro dogma, es el *reductivismo*, la creencia en que todo enunciado que tenga sentido es equivalente a alguna construcción lógica basada en términos que refieren a la experiencia inmediata. (Quine, 1951: 61)

Así es como Quine comienza el texto en el que rechaza las dos tesis a las cuales llama 'dogmas': la distinción analítico-sintético y el reduccionismo. La tarea de esta sección será pensar cómo tal discusión y tal postura pueden motivar al externismo. Esta será una tarea distinta y más modesta de la que realiza Davidson (2003) cuando encuentra en el Quine de "Two Dogmas of Empiricism" y de *Word and Object* "una forma potente de externismo" (Davidson, 2003: 281). La estrategia no será identificar al propio Quine como un externista –aunque Davidson tenga buenos argumentos en esta dirección– sino identificar cómo el rechazo de los dos dogmas puede favorecer una concepción externista de la mente.[8]

---

7 Esta interpretación es favorecida y utilizada en el último capítulo de esta disertación para explicar cuestiones referentes a auto-conocimiento, donde también se trata el enfoque expresivista con más atención.

8 La razón de mi elección por la vía más modesta es que la visión quineana sobre lo mental exigiría una discusión más detallada acerca de su adecuación al externismo. El conductismo que frecuentemente le es atribuido parece más bien marcar un escepticismo acerca de la propia esfera de lo mental (Quesada, 1987: 158). Quine (1975: 87) distingue tres niveles de explicación: el mental, el comportamental y el fisiológico. Entre los tres, considera el mental como el nivel más superficial que "apenas merece el nombre de explicación. El fisiológico es el más profundo y más

El enunciado "todo soltero es un no-casado" (o, lo que sería más usual en español, "ningún soltero está casado") se ofrece frecuentemente como ejemplo de enunciado analítico. Y la explicación que se da es que tal enunciado es analítico porque el significado de 'no-casado' está ya presente en el significado de 'soltero'. Tales tipos de enunciado solamente nos exigirían que comprendiéramos bien los términos que se están usando para decidir con respecto a sus valores de verdad. En contraste con los enunciados supuestamente sintéticos, no necesitaríamos ir al mundo para verificar si todos los solteros son, de hecho, personas que no se han casado. No sería necesario nada más que mirar al propio enunciado para llegar a establecer su valor de verdad. A su vez, los enunciados supuestamente sintéticos, como por ejemplo "la nieve es blanca", nos exigirían tener en cuenta más factores que el enunciado mismo; nos exigirían tener

ambicioso, y es el lugar para explicaciones causales" (Quine, 1975: 87). Pero, para los propósitos de teorizar sobre el lenguaje y la mente, el nivel comportamental es el más útil. Dice más: "no identificaría la mente completamente con disposiciones verbales; así como Ryle y Sellars, yo identificaría la mente con disposiciones comportamentales, siendo la *mayoría* verbales. Y entonces, habiendo construido a su vez disposiciones comportamentales como estados fisiológicos, llego a la así llamada teoría de identidad de la mente: los estados mentales son estados del cuerpo" (Quine, 1975: 94). Ver Quine (1985) y Gibson Jr. (2006: 196-9) para más sobre su teoría de la mente. Quine, al rechazar que los estados mentales son elementos de la psicología interna de un sujeto se aleja claramente del internismo, pero esto no lo localiza automáticamente en el externismo. Además, si los estados físicos con los cuales identifica los estados mentales son entidades como estados del sistema nervioso y éstos son suficientes para identificar un supuesto estado mental, Quine estaría más bien defendiendo un internismo. Quine parece dar elementos en las dos direcciones, cuando dice por ejemplo "el único cambio es que nosotros reconocemos estados mentales como estados del cuerpo en lugar de estados de otra sustancia, la mente" (Quine, 1985: 5) y "[u]n estado mental no se manifiesta siempre en la conducta. Físicamente analizado, es un estado del sistema nervioso. Podemos decir qué estado es y distinguir uno de otro, sin embargo, sin conocer el mecanismo neuronal. Lo especificamos con la ayuda de un término mental, que a su vez se aprendió por medio de señales comportamentales" (Quine, 1985: 6).

en consideración cosas como, por ejemplo, el color de la nieve, que según tal razonamiento, no componen el significado de la palabra 'nieve'. La noción de verdad analítica se puede entender al menos de dos formas distintas: i. el predicado de una oración analítica está contenido en el sujeto (por ejemplo, *soltero* contiene *no casado*); ii. una verdad analítica es una oración que es verdadera solamente en virtud de su supuesto significado, con independencia de cómo es el mundo. La argumentación de Quine ofrece bases para rechazar los dos sentidos, pero es el segundo el que más explícitamente puede vincularse al externismo.[9]

La principal crítica de Quine al primer dogma es la ausencia de criterios para distinguir entre enunciados analíticos y sintéticos. Según lo ve, cualquier intento de demarcación y de esclarecimiento del término 'analítico' sería insatisfactorio porque tendría que apelar a otras nociones igualmente necesitadas de esclarecimiento. Según Grice y Strawson (1956), clásicos críticos del trabajo de Quine, su estrategia podría resumirse de la siguiente manera: si un elemento que compone una familia de expresiones puede ser satisfactoriamente entendido o explicado, entonces los otros elementos podrían ser satisfactoriamente explicados con relación a tal término. En "Dos Dogmas del Empirismo", los términos con los que se intenta explicar el término 'analítico' son: 'auto-contradictorio', 'necesario', 'sinónimo', 'regla semántica' y 'definición'. Sin embargo, la conclusión de Quine es que cada elemento de la familia es necesario para la explicación de cualquier otro (Grice & Strawson, 1956: 147).

Quine, sin embargo, más que argumentar que la empresa de distinguir entre enunciados analíticos y sintéticos es fallida, denuncia que la propia motivación para querer sostener una concepción del lenguaje en términos de dos componentes sería filosóficamente

9 Preti (1995) también encuentra una relación entre el rechazo del primer dogma y el externismo, pero en el camino inverso al defendido aquí. Le parece que el propio externismo puede ser una motivación para rechazar el primer dogma.

problemática; una motivación que presupone la existencia de dos momentos: uno en el que el lenguaje actuaría en solitario y otro en el que sería el mundo el que actuara.

Tal suposición parece chocar con una posición externista. El primer dogma, tal y como lo presenta Quine, es en realidad consistente con la negación de la tesis externista expuesta inicialmente. El primer dogma permite que, por lo menos, algunos de los contenidos mentales (los que corresponden a enunciados analíticos) puedan ser individuados con independencia del mundo. Con el rechazo se abre la posibilidad de que todos los enunciados dependan del mundo y eso permite llegar a la idea de que es necesario referirse a factores externos a la piel del sujeto para identificar sus contenidos mentales. Habría, sin embargo, otras tres posibilidades de entender cómo se da esta dependencia y qué involucran los dos dogmas que no favorecería un externismo. La primera posibilidad, llamémosla 'híper-internismo', sería considerar la dependencia con el mundo como ya de por sí internista, como en el caso de teorías que encuentran en los *sense data* el vínculo entre nuestra mente y el mundo. Sí el mundo es un constructo lógico a partir de tales inputs, tanto enunciados analíticos como enunciados sintéticos son desde el principio individuados de una manera internista. Los inputs –que luego serán la base de todo el sistema de creencias– están más bien *en* la piel del sujeto (o alternativamente, en la cabeza del sujeto, como sensaciones privadas), por lo que sostener la distinción entre tales enunciados o rechazarla no dice nada con respecto a un resultado externista. Sin embargo, Quine argumenta en contra de esta imagen del mundo como un constructo lógico, pero sobre todo en contra del reduccionismo –el segundo dogma que discutiré a continuación– que en compañía del primer dogma, está en la base del híper-internismo.

La segunda posibilidad de entender el primer dogma sería entender los enunciados analíticos como establecidos en una esfera pública y por lo tanto, dependiente de un factor externo, a saber, el

social. Llamemos esta posibilidad 'internismo social'. En principio, parece que no estaríamos pasando de una imagen internista a una externista con el rechazo de la distinción analítico-sintético. Sin embargo, apelar solamente a la comunidad parece ser también un caso de internismo[10], que separa lo interno a la comunidad de mentes y lo externo a ella. Los enunciados analíticos entendidos de esta manera también carecerían del componente fáctico, lo que favorecería una idea internista, al ser posible identificar el contenido de un enunciado analítico solamente con referencia a las creencias disponibles a un sujeto, o incluso a lo disponible a la sociedad. Y en este sentido la crítica a la distinción analítico-sintético no se aplicaría solamente a una concepción del significado como constructo a partir de experiencias privadas, sino que también sería válida contra la idea de que las verdades analíticas hacen explícitas convenciones semánticas dentro de la comunidad. Es por eso que una vez rechazada la distinción sí parece haber un paso en dirección al externismo. Si uno rechaza la distinción analítico-sintético, y por lo tanto que las llamadas convenciones sociales o definiciones son establecidos sin interferencia del mundo, queda más claro ver cómo en determinadas circunstancias una definición también se modifica por medio de nuevos hallazgos empíricos. Un caso célebre es el del término 'átomo', que aunque se define inicialmente como la partícula más pequeña de la materia, ve cómo su significado se modifica por el descubrimiento de sus sub-partículas. Si consideramos que tener

10 En un sentido, lo que llamo 'internismo social' sería una especie de externismo social puro. Sin embargo, como he argumentado en el capítulo anterior, la distinción entre externismo social y externismo físico no es una buena distinción porque en general se sostiene que ambos factores son necesarios para la individuación de contenidos mentales. La propia posición de Burge, que es frecuentemente considerada como la representante del externismo social, no podría ser identificada como siendo en realidad un internismo social. Para Burge la relación con el mundo físico es primordial y la dependencia social no está delineada en términos de convenciones. También en el capítulo anterior argumento que es erróneo clasificar a Burge como externista social.

estados mentales con el concepto 'átomo' también depende del ambiente de esta manera, no es tan fácil aceptar que haya contenidos y estados mentales sin contribución del mundo.

Una vez que el rechazo de los dos dogmas lleva a la conclusión de que los significados (y los contenidos mentales) no pueden ser identificados por factores puramente internos a uno o a la comunidad, cabría una tercera posibilidad que sería el nihilismo con respecto al proprio significado. Llamemos a esta posibilidad 'nihilismo'. Esta parece ser de hecho la opción que Quine abraza [Fodor y LePore (1992), por ejemplo, defienden esta lectura de Quine], que combina un holismo confirmacional y un nihilismo semántico, pero no es necesario que se siga el nihilismo del rechazo de los dos dogmas.

La tesis del reduccionismo se refiere a la idea de que el significado sea reducible a algunas de sus partes, los átomos últimos (los inputs sensoriales, por ejemplo), o de una manera más específica, que las oraciones empíricas pudieran ser reducidas por definición a inputs sensoriales. En principio, el reduccionismo sería compatible con la idea de que determinados contenidos, los átomos últimos, se individuarían dependiendo del mundo. Pero como hemos visto anteriormente, la dependencia del mundo no parece implicar necesariamente ni un internismo ni un externismo. En el caso del híper-internismo, aunque los inputs sensoriales conecten el sujeto al mundo, son individuados de manera internista. Sin embargo, incluso en el caso de que tales inputs sensoriales fuesen individuados de manera externista, y en principio pareciera favorecer al externismo, el caso es que el dogma del reduccionismo es dependiente del dogma anterior. El reduccionismo funciona si existen situaciones en las cuales un enunciado "resulta confirmado vacuamente, *ipso facto*, ocurra lo que ocurra; [y] ese enunciado es analítico" (Quine, 1951: 85). El dogma del reduccionismo depende del primer dogma porque supone que el átomo último no tiene contribución alguna del sistema conceptual del sujeto (o del sistema de convenciones de la comunidad), ya que la contribución dependería exclusivamente del contenido no

conceptual de la experiencia del sujeto (o del mundo)[11]. Esto es ya afirmar que pueden existir enunciados con contribución nula de una de las partes. En realidad, dado que el rechazo del reduccionismo da lugar al holismo, se va también una importante razón para sostener la distinción analítico-sintético. Rechazar el reduccionismo es también rechazar "la suposición de que todo enunciado, aislado de sus compañeros, puede tener confirmación o invalidación" (Quine, 1951: 85). El holismo confirmacional de Quine involucra la idea que el papel de la comunidad aparece ya mezclado con el papel que juega el mundo:

> La totalidad de lo que llamamos nuestro conocimiento, o creencias, desde las más casuales cuestiones de la geografía y la historia hasta las más profundas leyes de la física atómica o incluso de la matemática o de la lógica puras, es un tejido de elaboración humana y que no está en contacto con la experiencia más que a lo largo de sus lados. (Quine, 1951: 86)[12]

Según Churchland (1979: 49) la consecuencia de desarrollar la idea del holismo confirmacional sería aceptar que cualquier oración podría ser tomada como verdadera, si estuviéramos dispuestos a pagar el precio de aumentar la complejidad y disminuir la coherencia del resto de nuestras creencias[13]. Visto desde esta perspectiva, no se

11 No es accidental que esta forma de plantear la relación entre los dos dogmas rechazados por Quine, tanto en su versión híper-internista como en su internista social, tenga ya ecos del rechazo por parte de Davidson del tercer dogma, que discutiré más adelante.

12 He modificado parcialmente la traducción de Manuel Sacristán.

13 Esto no nos impide reconocer que estamos dispuestos a cambiar el valor de verdad de determinados enunciados con mucha más facilidad que el de otros. De hecho, la comunicación parece depender de que algunas cosas estén fijas para que podamos plantearnos cambiar otras. El problema se encuentra en identificar los enunciados que se mantendrían más inmóviles con aquellos que dan los significados del lenguaje. Parece que lo más sensato sería tomarse la importancia semántica de una oración como una cuestión de grado, y es en ese sentido en el que el argumento quineano apunta. Wittgenstein, en *On Certainty* (1969), es una buena referencia para defender que, en las prácticas lingüísticas, es necesario

puede empezar eligiendo cómo vamos a llamar a las cosas del mundo para después desarrollar afirmaciones sobre él y es en este sentido que rechazar el dogma del reduccionismo (y abrazar el holismo) va junto con el rechazo de la distinción analítico-sintética.

El rechazo de los dos dogmas ofrece tanto razones para deshacerse del híper-internismo como del internismo social. Por un lado, una vez que se argumenta contra el reduccionismo, se diluye un elemento fundamental para hablar del mundo como un constructo lógico a partir de las experiencias empíricas. Por otro lado, si el mundo no es un constructo lógico, al rechazar la distinción analítico-sintético, los factores externos al sujeto pasan a poblar sin discriminaciones su sistema de creencias. Sin embargo, tras el rechazo de los dos dogmas, parece que serían posibles otras dos alternativas: el nihilismo y externismo. La primera opción, supuestamente favorecida por Quine se compone por el holismo confirmacional acompañado por el nihilismo semántico, mientras que en el camino externista, el holismo confirmacional viene acompañado por el holismo semántico.

Como hemos visto, el dogma del reduccionismo y la distinción analítico-sintético permiten concebir nuestro lenguaje como estando compuesto por elementos con contribución exclusiva de la comunidad y elementos con contribución exclusiva del mundo. Los *sense data* serían ejemplos de los últimos. Por eso, rechazar los dos dogmas pone en cuestión la propia inteligibilidad de algo como los *sense data* que prescinde de elementos lingüísticos. Sin embargo, que no podamos separar la contribución del mundo de la contribución lingüística (o que no podamos separar datos de convenciones), no implica que los dos tipos de contribuciones no sigan siendo intrínsecamente distintos. Y esta conclusión débil del rechazo de

mantener siempre algo fijo, cuestión puesta de manifiesto a través de su analogía con las bisagras de las puertas. Lo que Wittgenstein nos enseña es que las bisagras no pueden estar fijas por sí mismas; cuando nos ponemos a pensar sobre ellas, son necesarias otras bisagras.

los dos dogmas parece ser compatible todavía con la distinción analítico-sintético en una versión holista: ahora dentro del lenguaje, pensamiento o ciencia como un todo. Es decir, si uno no sigue la conclusión quineana nihilista (que de una cierta manera también se basa en la distinción entre los tipos de contribuciones), le quedaría como opción el externismo de dos factores, en el sentido de que nuestros sistemas de creencias, como un todo, estuvieran poblados por elementos internos y externos a uno o a una comunidad[14]. Sin embargo, uno podría insistir en rechazar también esta forma holista de la distinción analítico-sintético, defendiendo que no solamente no podemos distinguir entre contribuciones lingüísticas y del mundo, sino que sencillamente no hay dicha distinción. Es decir, la conclusión de que no podemos demarcar dónde comienza y dónde termina el papel de la comunidad en la individuación de contenidos respondería más bien a una actitud positiva externista que insiste que elementos que no tengan forma proposicional no pueden interferir en nuestros sistemas de creencias. Algo como los *sense data* es ininteligible porque es ininteligible concebir que factores del mundo puedan afectar nuestros sistemas de creencia como un todo, sin que ya tengan ellos mismos una forma proposicional.

Estas conclusiones seguramente no son las de Quine, sino las de Davidson, que denuncia además de los dos dogmas, el tercero, que en este contexto puede caracterizarse como la disolución del contraste entre la contribución del mundo y de la comunidad. Sin embargo, si hemos llegado a esta conclusión solamente rechazando la distinción analítico-sintético en su versión más fuerte (que he llamado 'holista'), parece ser que la posición de Quine es inestable en el sentido de tender a una posición como la de Davidson. Es decir, parece que rechazar los dos dogmas tiende al rechazo del

14 Esta sería una variación de la formulación del externismo de dos factores desarrollado en el capítulo pasado, en términos de que cada estado mental estuviera compuesto por un elemento estrecho y otro amplio.

tercero, si uno lleva a las últimas consecuencias el rechazo del dualismo analítico-sintético. Eso claramente le costaría a Quine su empirismo (como Davidson constata), pero rechazar los dos dogmas parece tener el efecto demoledor que encuentra Davidson. Y en este sentido, el externismo de dos factores que surge con la versión tradicional del rechazo de los dos dogmas, también parece tender a un externismo global, como el que surge con Davidson y que será discutido a continuación.

3. Argumento contra el Tercer Dogma (AcTD) – Donald Davidson

El llamado 'tercer dogma', rechazado por Davidson, se refiere a la distinción entre esquema conceptual y contenido empírico. El camino emprendido por Davidson es, como él mismo sugiere (1974), una radicalización de la conclusión quineana lograda con el rechazo de los dos primeros dogmas del empirismo. Los datos no pueden separarse de las convenciones. La tarea aquí será pensar cómo esa maniobra puede vincularse a una motivación externista. Al igual que en la sección anterior, el interés estará más en ver cómo este paso en concreto puede servirnos para defender el externismo, que en hacer exégesis del externismo propiamente davidsoniano[15].

El dualismo en cuestión mantiene que los esquemas conceptuales serían "maneras de organizar la experiencia; sistemas de categorías que darían forma a los datos de la sensación; o incluso los puntos de vista a partir de los cuales los individuos, culturas o periodos harían el inventario de lo que son testigos" (Davidson, 1974: 183). El dualismo entre esquema conceptual y contenido forma parte, por tanto, de una manera de pensar sobre nuestras capacidades cognitivas y lingüísticas como actividades organizadoras

15 El externismo davidsoniano es discutido en el próximo capítulo de manera más detallada.

de una materia bruta ofrecida por el mundo por medio de nuestras sensaciones. Recusar tal distinción es recusar esta imagen, una imagen que está compuesta por dos dualismos: la separación entre conceptos y sensaciones desnudas, y la separación entre visiones del mundo y el mundo. Como señala Pinedo (2004: 271), rechazar el tercer dogma significa rechazar ambos dualismos.

De acuerdo con Davidson (1974: 198), el tercer dogma podría servir fácilmente de base para llegar al relativismo conceptual y a la noción de verdad relativa a un esquema. El dualismo esquema conceptual y contenido presupondría ya la imagen de varios esquemas conceptuales que comparten un contenido común, posiblemente intraducibles entre sí. Sin embargo, la idea de un esquema conceptual alternativo es, según Davidson una noción ininteligible. Esa será la principal acusación que Davidson dirija al tercer dogma: ha de ser rechazado porque da lugar al relativismo de esquemas conceptuales, una idea que a su vez da lugar a una noción ininteligible, la de "esquema conceptual alternativo". Davidson defiende que desde que se reconoce a alguien como poseedor de un esquema conceptual se deja de estar en posición de juzgar que tal ser tiene conceptos o creencias radicalmente diferentes de las nuestras (Davidson, 1974: 197). Es decir, a quien está en una situación de interpretación radical (como, por ejemplo, en el famoso caso propuesto por Quine de la traductora que se va a una comunidad aislada de cualquier otra) se le exige que sea capaz de identificar qué sonidos cuentan cómo comportamientos lingüísticos y cuáles no. Y esto descarta la posibilidad del aislamiento entre esquemas. Reconocer algo como "alguien" es de por sí una tarea nada despreciable. Es estar ya en una determinada posición de comprensión de aquel sistema de conceptos, de forma que éste no podría simplemente estar aislado del mi supuesto sistema conceptual (ver Davidson 1974). La consecuencia de esa posición con respecto a la interpretación es que no es posible percibir a alguien como un ser lingüístico y, a la vez, juzgarle como usuario de una lengua ininteligible

(y pensar que tiene una visión completamente distinta de la mía).[16] Davidson defiende que significado y mentalidad mantienen una relación de interdependencia que responde a la interdependencia de dos aspectos de la interpretación de los enunciados: la atribución de creencias y la interpretación de los enunciados (Davidson, 1974: 195); dos aspectos de la interpretación que ocurren a la vez. Esta línea de razonamiento compone la primera parte del argumento davidsoniano contra el tercer dogma, que puede reestructurarse de la siguiente manera:

1. La imposibilidad de intertraducibilidad es condición necesaria para la diferencia entre esquemas conceptuales (según la definición de 'esquemas conceptuales diferentes' que utiliza Davidson);
2. No puede haber casos de imposibilidad, completa o parcial, de intertraducibilidad (por la argumentación delineada anteriormente);

CONCLUSIÓN 1: No podemos hablar de diferentes esquemas conceptuales.

Davidson entiende que Quine ha dado pasos importantes, pero no suficientes, para lograr lo que se buscaba. Rechazar solamente los dos primeros dogmas sería todavía compatible con la posibilidad de que una sociedad pudiera cambiar completamente los parámetros de su sistema, hasta el punto de surgir esquemas intraducibles en relación a los antiguos. Y esto significa que podríamos concebir todavía el mundo de un lado y la mente de otro. Según Davidson lo que se buscaba era garantizar el contacto con el mundo y para alcanzarlo es necesario rechazar también el tercer dogma.

El tercer dogma, sin embargo, no se refiere solamente a la idea de que haya un mundo, de un lado, y diferentes esquemas, del otro, que organizan tal materia bruta en diferentes sistemas. El

16 Ver Borgoni (2006) para una argumentación acerca de las consecuencias de las tesis davidsonianas para la diversidad cultural.

tercer dogma se refiere a la idea misma de que haya esta división, que incluye la posibilidad de que hubiera un solo esquema conceptual compartido por todos nosotros que contrastara con el mundo. El argumento davidsoniano completo, por lo tanto, tiene la siguiente estructura:

1'. Para que haya distinción entre esquema y contenido es necesario o bien una diferencia entre esquemas o bien un sólo esquema que contraste con el mundo;
2'. No hay diferentes esquemas (conclusión 1).
3'. No hay un solo esquema que contraste con el mundo.
CONCLUSIÓN 1'. No hay distinción esquema-contenido.

La premisa 3' es también defendida por Davidson con referencia a las condiciones de la interpretación. Davidson defiende que la interpretación es posible porque compartimos gran parte de nuestras creencias y, además, al interpretar a alguien le atribuimos un gran grado de corrección. Este paso es parte del llamado 'principio de caridad', que se refiere en parte al hecho de que interpretar a alguien nos exige que dejemos algunas cosas estables: no podemos al mismo tiempo preguntarnos por el significado de una palabra, cuestionar si el interpretado la está usando bien y dudar si tal palabra quiere decir algo. Si Davidson defiende que el único modo de interpretarnos es presuponer el acierto[17], también defiende que muchas de nuestras creencias, que son compartidas, son de hecho verdaderas[18], por la naturaleza interpersonal de la objetividad. Este

17 Davidson (1974: 197) insiste en que aplicar el principio de caridad no es una opción, sino una condición de la interpretación.

18 Hay quienes mantienen que el argumento davidsoniano en favor de que tengamos conocimiento se resume al así llamado 'argumento del intérprete omnisciente'. No comparto de esta visión. Al igual que Manning (1995: 346), pienso que una vez que se tienen en consideración las condiciones de la "interpretación radical", dicho argumento no añade nada a los demás argumentos de Davidson. Yo añadiría que los argumentos involucrados en la interpretación radical hacen ininteligible la idea misma de un intérprete omnisciente. Es decir, tanto el error masivo como el conocimiento masivo son ininteligibles desde la

paso señala al conocimiento como otra condición de posibilidad de la interpretación y, por lo tanto, del propio contenido mental, dado que la mente no es nada más allá de lo atribuido y detectado en la interpretación.

Esta es la imagen externista que se puede lograr con el rechazo del tercer dogma: la mente –o visión de mundo– es solamente concebible como tal si el mundo está presente desde el principio. Que nos relacionemos lingüísticamente es para Davidson evidencia de que nuestras creencias no están despegadas del mundo. Y estar pegado al

---

perspectiva de la interpretación radical. En realidad, el mejor candidato a ser el intérprete omnisciente sería el propio escéptico (Marton, 1999: 77), el ser que por dudar de todo no podría ser un intérprete real. Parece que solamente un ser que está situado siempre en la perspectiva de tercera persona, es decir, fuera de las relaciones interpersonales, podría ser o bien el ser completamente equivocado o el ser completamente correcto.

Davidson desarrolla el argumento del intérprete omnisciente en "The Method of Truth in Metaphysics" (1977) y en "A Coherence Theory of Truth and Knowledge" (1983). Nos pide que imaginemos un ser omnisciente (sobre el mundo) interpretando a los demás seres falibles. Este intérprete especial usa el mismo método de interpretación que las personas falibles usan: atribuye creencias a los demás e interpreta sus discursos sobre la base de sus propias creencias. Haciendo esto, el intérprete omnisciente encuentra a aquellos que interpreta ampliamente consistentes y correctos. Davidson concluye: "está claro porqué el error masivo sobre el mundo es sencillamente ininteligible, porque suponer que esto sea inteligible es suponer que podría haber un intérprete (el omnisciente) que interpretara correctamente a alguien como estando equivocado masivamente, y hemos mostrado que esto es imposible" (Davidson, 1977: 201). Se pueden encontrar críticas a este argumento en Foley y Fumerton (1985) y Dalmiya (1990). Foley y Fumerton (1985) por ejemplo, acusan Davidson de equivocarse en intentar alcanzar la realidad del mundo externo desde la posibilidad de la existencia del intérprete omnisciente. Dalmiya (1990) ofrece un contraejemplo al caso de Davidson imaginando ahora un intérprete engañado (*deluded*) que en lugar de compartir creencias verdaderas, comparte solamente las falsas. Estoy de acuerdo que el caso del intérprete omnisciente no es una buena instancia para establecer que somos conocedores. Sin embargo, la conclusión de que somos conocedores la logra Davidson por su visión general sobre la interpretación y su tesis sobre la interdependencia entre los tipos de conocimiento (Davidson, 1991), y esto hace innecesaria cualquier referencia al argumento del intérprete omnisciente.

mundo significa tener conocimiento, tener un estado mental que no puede prescindir del carácter externo. Según he defendido en el capítulo pasado, esta conclusión es externista en el sentido de lo que he llamado 'externismo constitutivo': los contenidos (y estados) mentales deben ser individuados, en parte, por factores externos a la piel del sujeto porque la mente está constituida por conocimiento.

El resultado externista es más evidente con relación al rechazo del dualismo esquema-contenido entendido en términos del divorcio entre visiones del mundo y el propio mundo. Tras el rechazo del dogma, es posible afirmar que una visión del mundo conlleva una relación intrínseca con el mundo, en este caso, en términos de conocimiento del mismo. Sin embargo, el otro sentido del dogma en términos de la separación entre conceptos y sensaciones desnudas también nos dice algo sobre la individuación de estados mentales. Su rechazo significa que sólo en un ambiente ya lingüístico, podríamos pensar en la individuación de contenidos mentales. De hecho, McDowell encuentra un paralelo entre el dualismo esquema-contenido y el argumento wittgensteiniano contra el lenguaje privado en los siguientes términos: "La idea es que el 'lingüista privado' sucumbe a una versión del dualismo entre esquema y dado: su idea […] es que el flujo de la conciencia está constituido por elementos no-conceptuales que justifican conceptualizaciones suyas" (McDowell, 1998: 280). El lema "solamente las creencias justifican creencias" (ver Davidson 1983), que es una de las conclusiones importantes de la posición de Davidson, adquiere el sentido de que la inmersión en las prácticas de una comunidad lingüística es el único recurso para hablar sobre contenidos y, de este modo, individuarlos. Pero, dado que una comunidad lingüística es evidencia de que el mundo viene unido a ella, el peso puesto en la esfera conceptual no se contrapone al mundo propiamente dicho, del cual somos parte.[19]

19 Para McDowell (1994, conferencia 1), el rechazo del tercer dogma bajo la lectura de que elementos no conceptuales no pueden justificar elementos conceptuales,

4. Rechazo del Máximo Factor Común (MFC) – John McDowell

> Una tentación que apunta de forma todavía más profunda en la dirección de la concepción del "Máximo Factor Común" puede expresarse así: "*Ex hyphotesi* una simple apariencia puede ser indistinguible de lo que describes como un hecho manifiesto. Así, en un determinado caso, uno no puede decir con seguridad si lo que alguien afronta es uno u otro de tales casos ¿Cómo entonces puede haber una diferencia en lo que es dado en la experiencia, en cualquier sentido que pudiera tener importancia para la epistemología?" (McDowell, 1982: 389-90)

La noción del máximo factor común (MFC) contra la cual argumenta J. McDowell responde a la suposición de que hay algo común entre la experiencia real y la mera apariencia. Según lo entiende McDowell, tal elemento tiene que ver con una cierta forma de ver la experiencia, concebida en términos de satisfacción de criterios, que se compromete con la idea de un intermediario entre la mente y el mundo. Bajo dicha concepción, en una experiencia de ver algo rojo, por ejemplo, tendríamos tres elementos: el objeto rojo en el mundo, la imagen de tal objeto rojo en mi mente y mi mente teniendo acceso a esa imagen mental. En una alucinación de un objeto rojo, podríamos tener la misma imagen mental de un objeto rojo ante mi mente. La indiscernibilidad entre las dos imágenes mentales a las cuales mi mente tendría acceso no me permitiría saber de forma inequívoca en cuál de las situaciones me encuentro.

La noción de MFC rechazada por McDowell esconde en realidad un paquete de nociones que son revisadas en conjunto: la concepción de la experiencia que se compromete con un

---

que es traducida por el lema davidsoniano de que "solamente creencias justifican creencias", resulta en un coherentismo sin anclaje en el mundo. Davidson, sin embargo, puede esquivar esta crítica enfatizando que una comunidad lingüística no es solamente coherente, sino que además tiene conocimiento, en la línea argumental expuesta anteriormente.

intermediario entre la mente y el mundo; la idea de acceso mental a objetos internos a la mente; y una concepción internista de criterios de atribución de conocimiento[20]. La argumentación de McDowell se basa especialmente en apuntar el problema que plantea tales nociones, esforzándose en dejar claro qué otras alternativas serían más fértiles.

Lo que interesa a McDowell en su texto (1982) es pensar acerca de las experiencias no solamente del mundo exterior sino también de las propias vivencias subjetivas. El externismo en la forma expuesta tradicionalmente por medio del experimento mental de Putnam (1975) concluye que los contenidos de los enunciados acerca del agua que profieren Oscar1 y Oscar2 son distintos porque sus ambientes son distintos. El significado del enunciado de Oscar1 respondería a la constitución H_2O del agua en la Tierra y el enunciado de Oscar2 a la constitución XYZ del agua en la Tierra Gemela. La conclusión externista en lo que respecta a la semántica distingue las dos

20 Sawyer y Mayors (2005) examinan la posibilidad de defender el externismo de lo mental vía el externismo epistémico. Ellos defienden que el externismo epistémico depende del propio externismo sobre el contenido mental para sostener una explicación satisfactoria acerca de la confiabilidad en la formación de creencias: "solamente apelando al externismo sobre el contenido uno puede ofrecer una restricción de principio y efectiva sobre el conjunto de mundos con respecto a los cuales uno sería confiable con el fin de que las creencias perceptuales de uno estuvieran justificadas" (Sawyer & Mayors, 1995: 281). Sin embargo, el propio externismo epistémico serviría como apoyo al externismo sobre el contenido. Uno de sus objetivos explícitos es buscar una ruta hacia el externismo del contenido que no requiera consideraciones acerca de experimentos mentales del tipo tierra gemela (Sawyer & Mayors, 1995: 257). En un cierto sentido, la crítica de McDowell al máximo factor común involucra consideraciones epistémicas. Si este movimiento de McDowell sirve para sostener un externismo de lo mental, mi argumentación se adecuaría al proyecto de Sawyer y Mayors (2005). Sin embargo, no discuto en este trabajo la relación entre externismo epistémico y externismo sobre lo mental en términos generales. Para referencias clásicas sobre externismo epistémico naturalistas, ver Quine (1969) y Goldman (1979, 1994) y para versiones no naturalistas (o menos naturalistas) del externismo epistémico ver Davidson (1990) y Brandom (2000).

situaciones claramente. Pero no lo hace con respecto a la experiencia que los dos Oscar tienen frente al líquido que llaman agua. Por lo que allí se puede entender, las experiencias serían idénticas. Y la razón más inmediata para tal posición sería decir que la apariencia de XYZ es indistinguible de la apariencia de H_2O. De hecho, Putnam propone su experimento mental en términos similares a este. Las conclusiones mcdowellianas nos permitirán ir más allá y afirmar la distinción de las dos situaciones en los dos niveles.

La concepción de la experiencia como satisfacción de criterios es la concepción acusada por McDowell de ser la responsable de impedirnos concebir las dos situaciones como totalmente distintas. Parece que podríamos llegar más lejos que en la distinción trazada por Putnam. Si tener la experiencia de ver al XYZ es satisfacer criterios como los que la apariencia del agua proporciona, entonces la misma posibilidad podría ocurrir en las dos tierras. La cuestión es que los criterios son falibles, y experimentar la satisfacción de un criterio se vuelve enteramente compatible con la situación en la que el enunciado sea falso. Y esto sería incoherente.

El mismo razonamiento se aplica a la atribución de estados mentales a nuestros pares. Si lo concebimos en términos de satisfacción de criterios, podríamos decir que un comportamiento de dolor y su imitación serían indiscernibles al satisfacer ambos los mismos criterios de apariencia de tener dolor. Pero "esto nos llevaría a la siguiente tesis: saber que alguien está en algún estado interno puede estar constituido por estar en una posición en la cual, por lo que uno sabe, la persona puede no estar en ese estado 'interno'. [Y] esto parece ser francamente incoherente" (McDowell, 1982: 371).

En lugar de tal noción de experiencia, McDowell propone la noción disyuntiva, que concibe a las situaciones reales y las de ilusión, desde el inicio, como situaciones distintas entre sí. Según McDowell, frente a una ilusión, lo que se experimenta es la *apariencia* del agua, lo que se distingue de la experiencia del agua misma. Como ha sido señalado, McDowell indica que no aceptar tal

distinción sólo tendría una razón: el compromiso con la existencia de un intermediario entre la mente y el mundo. El intermediario que posibilitaría la compatibilidad entre las situaciones de acierto y de error. Una vez abandonada tal imagen, el MFC pierde su fuerza, posibilitando un contacto directo entre mente y mundo, que es la opción mcdowelliana más significativa. Según lo entiende, si uno adopta la concepción disyuntiva de las apariencias, tendría que tomarse en serio la idea de una apertura de la experiencia a la realidad externa sin mediación del sujeto, mientras que el MFC nos permitirá diseñar una interfaz entre ellos (McDowell, 1982: 392).[21]

El rechazo del MFC que da lugar a la concepción disyuntiva de la experiencia tendría un fuerte atractivo externista, aunque es necesario notar que el rechazo del MFC no implica necesariamente dicha noción de experiencia. Según lo ve también Macarthur (2003: 179) el externismo se muestra más evidente porque "la posibilidad misma del contenido empírico depende del hecho de que algunas de nuestras experiencias deben ser no-engañosas, en el sentido de que objetos reales figuran en ellos"; sigue: "a menos que haya algunos casos de percepción verídica no habría manera de que tuviéramos pensamiento con contenido empírico". Es decir, los contenidos no pueden ser identificados solamente con respecto a factores internos a la mente de uno, porque en primer lugar, ya no hay una imagen mental a la cual adherirse para hablar de significado, y en segundo lugar, el mundo tiene que estar presente, por lo menos en aquellos

21 Vega (2006) sostiene, sin embargo, que la concepción disyuntiva de la experiencia parece estar en tensión con otras bases de la posición de McDowell, como por ejemplo su compromiso con la idea sellarsiana de que conocer es ocupar un lugar en el espacio de las razones (Sellars, 1956). El externismo resultante de la noción disyuntiva de McDowell parece adecuarse al que he identificado en el capítulo anterior como un 'externismo extrínseco', que explica la externalidad de la mente por referencia a relaciones causales. Se aplicáramos esta idea a la noción disyuntiva, podríamos entender que una experiencia real de ver una manzana debe su status al hecho de que hay alguna relación causal entre mi mirar y la manzana, mientras que en una alucinación, no hay tal relación.

casos de percepción verídica. La visión disyuntiva de la experiencia, por lo tanto, mantiene la posibilidad de la falibilidad "mientras preserva la idea de que en la experiencia estamos en contacto cognitivo con el mundo" (Vega, 2006: 68).

Dado este paso con el rechazo de la noción del MFC, podemos entender que una situación de ver algo rojo y de tener una alucinación de algo rojo no pueden tener nada en común, son no solamente semánticamente distintas sino también epistemológicamente. En la primera situación el sujeto conoce mientras en la otra, no. Sin embargo ¿cómo se le aparecen tales experiencias al propio sujeto?

Una de las razones detrás de la conservación de la noción de MFC era la indiscernibilidad entre la situación real y la alucinación, porque existía algo en común entre ambas. Tras el rechazo del MFC, ya no hay nada en común. ¿Se conquista así la certidumbre para el sujeto de las experiencias? Ciertamente no. La atribución de conocimiento o de equívoco a esos sujetos cognitivos no es garantía de que esos mismos individuos adquieran conocimiento sobre sí mismos en el sentido de saber en cuál de las situaciones están. Aceptar que un individuo posea conocimiento aunque no sepa que conoce es un paso que forma parte del rechazo del MFC.

McDowell dice que si no hay intermediario entre mente y mundo y por tanto, no se abre espacio para algo común a las dos situaciones, la de acierto y la de error, entonces estas habrán de ser caracterizadas siempre como situaciones distintas. Sin embargo, aceptar por completo la disyunción entre esas dos situaciones requiere que aceptemos que el conocimiento de en qué situación se encuentra uno no está garantizado. O sea, aceptar que una situación de experiencia real y una de ilusión son, de hecho, situaciones completamente dispares, no debe implicar que podamos distinguirlas nosotros mismos. Por esto, además de un cambio en la noción de experiencia, es necesario un cambio en los criterios de lo que se toma como conocimiento: dejamos de pensar que el sujeto cognitivo tiene acceso completo a sus estados mentales y que

estos estarían compuestos por imágenes mentales frente a las que la mente se encontraría para realizar cualquier actividad cognitiva. Por esto, el sujeto no necesita saber que sabe para que le atribuyamos conocimiento.

Conclusión

El internismo sostiene que la mente puede estar auto-contenida. Lo que los cuatro argumentos han hecho ha sido mostrar la incoherencia involucrada en tal idea desde diversas perspectivas. Wittgenstein, por ejemplo, ha señalado la incoherencia que conlleva la idea de lenguaje privado; Quine ha cuestionado la idea de que podamos tomar determinados términos de nuestro lenguaje como compuestos por factores puramente lingüísticos y otros puramente fácticos. Davidson, a su vez, ha mostrado que el problema está en que todavía podamos concebir la mente como algo separable del mundo, mientras que McDowell ha deconstruido la noción de máximo factor común.

He intentado mostrar que a ese nivel, todos los argumentos podían conducirnos hacia el externismo al volver incoherente la concepción internista de lo mental. Hemos presentado, sin embargo, algunas salidas positivas que abren paso a que de hecho podamos defender que los contenidos mentales *son* individuados por factores externos a la piel de uno. Con Wittgenstein, encontramos este aspecto cuando percibimos que los criterios de corrección de nuestro lenguaje sólo podrían ser concebidos como criterios externos no solamente al sujeto, sino externos a la comunidad. Con Davidson, hablar sobre mente y contenidos mentales ya trae consigo el mundo. Y con McDowell, hemos podido lograr otro aspecto externista con el rechazo del MFC, en términos de aceptar que nada respecto al conocimiento está garantizado, con la excepción de que somos conocedores.

Referencias Bibliográficas

Bishop, M. 1999, "Why Thought Experiments are Not Arguments", *Philosophy of Science* 66, pp. 534-41

Boghossian, P. 1989b, "The Rule-Following Considerations", *Mind* 98, pp. 507-49.

Bokulich, A. 2001, "Rethinking Thought Experiments", *Perspectives on Science* 9 (3), pp. 285-307.

Borgoni, C. 2006, "Davidson on Intercultural Dialog", en Gasser *et al.* (eds) 2006, *Cultures: Conflict – Analysis - Dialog* (29th International Wittgenstein Symposium), Kirchberg am Wechsel (Austria), Austrian Ludwig Wittgenstein Society, pp. 47-9.

Borgoni, C. 2009b, "En casa, en el mundo: el externismo global constitutivo", *Teorema* 23 (3), en prensa.

Brandom, R. 2000, *Articulating Reasons*, Cambridge, Mass., Harvard University Press.

Brendel, E. 2004, "Intuition Pumps and the Proper Use of Thought Experiments", *Dialectica* 58 (1), pp. 88-108.

Brown, J. R. 1991a, *The Laboratory of the Mind*, London/New York, Routledge.

Brown, J. R. 1991b, "Thought Experiments: A Platonic Account", en T. Horowitz and G. Massey (eds.), 1991, *Thought Experiments in Science and Philosophy*, Savage, MD, Rowman and Littlefield.

Burge, T. 1979, "Individualism and the Mental", en P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 21-83.

Churchland, P. 1979, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press.

Dalmiya, V. 1990, "Coherence, Truth and the 'Omniscient Interpreter'", *The Philosophical Quarterly* 40 (158), pp. 86-94.

Davidson, D. 1973, "Radical Interpretation", en D. Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 125-39.

Davidson, D. 1974, "On the Very Idea of a Conceptual Scheme", en D. Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 183-98.

Davidson, D. 1977, "The Method of Truth in Metaphysics", in D.

Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 199-214.

Davidson, D. 1983, "A Coherence Theory of truth and Knowledge", en D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 137-53.

Davidson, D. 1990, "Epistemology Externalized", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 193-204.

Davidson, D. 1991, "Three Varieties of Knowledge", D. Davidson, 2001 *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 205-20.

Davidson, D. 2003, "Quine's Externalism", *Grazer Philosophische Studien* 66, pp. 281-97.

Finkelstein, D. H. 2008, *Expression and the Inner*, Cambridge, Mass., Harvard University Press.

Fodor, J. & LePore, E. 1992, *Holism: A Shopper's Guide*, Cambridge, Mass., Blackwell Publishers.

Foley, R. & R. Fumerton 1985, "Davidson's Theism?" *Philosophical Studies* 48, pp. 83-89.

Gendler, T. 2000, *Thought Experiment: On the Powers and Limits of Imaginary Cases*, NY, Routledge.

Gendler, T. 2005, "Thought Experiments", en D. Borchert (ed.) 2006 *Encyclopedia of Philosophy* 9, Detroit, Macmillan Reference, pp. 388-94.

Gert, B. 1986, "Wittgenstein's Private Language Argument", *Synthese* 68, pp. 409-39.

Gibson Jr., R. F. 2006 "Quine's Behaviorism cum Empiricism" en R. Gibson (ed.) 2006, *The Cambridge Companion to Quine*, Cambridge, Cambridge University Press, pp.181-99.

Goldman, A. 1979, "What Is Justified Belief?", en Kornblith, H. (ed). 1994, *Naturalizing Epistemology*, Cambridge, Mass., The MIT Press, pp. 105-30.

Goldman, A. 1994, "Naturalistic Epistemology and Reliabilism", en French, P., Uehling, T. & H Wettstein. (eds.). *Midwest Studies in Philosophy 19: Philosophical Naturalism*, Notre Dame, University of Notre Dame Press, pp. 301-20.

Grice, H. P & Strawson, P. F. 1956, "In Defense of a Dogma", *The Philosophical Review* 65 (2), pp. 141-58.

Hacker, P. M. S. 1990, *Wittgenstein: Meaning and Mind*, Oxford, Basil Blackwell.

Horowitz, T. and Massey. G. (eds.), 1991, *Thought Experiments in Science and Philosophy*, Savage, MD, Rowman and Littlefield

Kripke, S. 1982, *Wittgenstein on Rules and Private Language*, Oxford, Basil Blackwell.

Kuhn, T. 1964, "A Function for Thought Experiments", T. Kuhn 1977, *The Essential Tension*, Chicago, University of Chicago Press, pp. 240-65.

Macarthur, D. 2003, "McDowell, Scepticism, and the 'Veil of Perception'", *Australasian Journal of Philosophy* 81 (2), pp. 175-90.

Manning, R. N. 1995, "Interpreting Davidson's Omniscient Interpreter", *Canadian Journal of Philosophy* 25 (3), pp. 335-74.

Marton, P. 1999, "Ordinary versus Super-omniscient Interpreters", *The Philosophical Quarterly* 49 (194), pp. 72-7.

McDowell, J. 1982, "Criteria, Defeasibility, and Knowledge", en J. McDowell 1998, *Meaning, Knowledge & Reality*, Cambridge, Harvard University Press, pp. 369-94.

McDowell, J. 1984a, "Wittgenstein on Following a Rule", en A. Miller & C. Wright (eds.) 2002, *Rule Following & Meaning*, Chesham, Acumen, pp. 45-80.

McDowell, J. 1994, *Mind and World*, Cambridge, Harvard University Press.

McDowell, J. 1998, "One Strand in the Private Language Argument" en J. McDowell 1998, *Mind, Value and Reality*, Cambridge, Mass., Harvard Univerity Press, pp. 279-296.

Miller, A. 2002, "Introduction", en A. Miller & C. Wright (eds.) 2002, *Rule Following & Meaning*, Chesham, Acumen, pp. 1-15.

Norton, J. 1996, "Are Thought Experiments Just What You Always Thought?" *Canadian Journal of Philosophy* 26, pp. 333-66.

Norton, J. 2004, "On Thought Experiments: Is There More to the Argument?" *Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association, Philosophy of Science* 71, pp. 1139-1151.

Pinedo, M. 2004, "The Anomalous Character of Experience", en J. C. Marek & M. Reicher (ed.) 2004, *Experience and Analysis* (Proceedings of the 27th International Wittgenstein Symposium), Kirchberg am Wechsel (Austria): Austrian Ludwig Wittgenstein Society, pp. 269-71.

Prades, J. L. 2006 "Varieties of Internal Relations: Intention, Expression

and Norms, *Teorema* 25 (1), pp. 137-154.
Preti, C. 1995, "Externalism and Analyticity", *Philosophical Studies* 79 (3), pp. 213-36.
Preti, C. 2002, "Normativity and Meaning: Kripke's Skeptical Paradox Reconsidered", *The Philosophical Forum* 33 (1), pp. 39-62.
Putnam, H. 1975, "The Meaning of 'Meaning", en H. Putnam 1975, *Mind, Language and Reality,* Philosophical Papers, vol. 2, Cambridge, Cambridge University Press, pp. 215-71.
Quesada, D. 1987, "Creencia, conducta y contexto", en J. J. Acero & T. Calvo (eds.) 1987, *Symposium Quine*, Granada, Universidad de Granada, pp. 157-74.
Quine, W.v.O. 1951, "Dos Dogmas del empirismo" (traducido por M. Sacristán), en *Desde un Punto de Vista Lógico*, Barcelona, Paidós, 2002, pp. 61-91.
Quine, W.v.O. 1960, *Word and Object*, Cambridge, Mass., MIT Press.
Quine, W.v.O. 1969, "Epistemology Naturalized", en Kornblith, H. (ed.) 1994, *Naturalizing Epistemology*, Cambridge, Mass., The MIT Press, pp. 15-31.
Quine, W.v.O. 1975, "Mind and Verbal Dispositions", en S. Guttenplan (ed.) 1977, *Mind and Language*, Oxford: Clarendon Press, pp. 83-95.
Quine, W.v.O. 1985, "States of Mind", *The Journal of Philosophy* 82 (1), pp. 5-8.
Sawyer, S. & Majors, B. 2005 "The Epistemological Argument for Content Externalism", *Philosophical Perspectives* 19, pp. 257-80.
Sellars, W. 1956, *Empiricism and the Philosophy of Mind*, reimpreso en 1997, Cambridge, Mass., Harvard University Press.
Vega, J. 2006, "Appearances and Disjunctions: Empirical Authority in McDowell's Space of Reasons", *Teorema* 25 (1), pp. 63-81.
Wittgenstein, L. 1953, *Investigaciones Filosóficas* (traducido por A. García Suárez & U. Moulines), Barcelona, Crítica, 1988.
Wittgenstein, L. 1969, *On Certainty* (traducido por G. E. M. Anscombe & D. Paul), Oxford, Basil Blackwell, 1979.

Summary of Chapter:

2 Externalism without Thought Experiments

ABSTRACT
The aim of this paper is to think about alternative ways of defending externalism that dispense with the use of thought experiments. Four lines of arguments will be studied: Wittgenstein's argument against private language and his considerations about following a rule; the Quinean arguments against the two dogmas of empiricism; the Davidsonian arguments against the third dogma; and, finally, the arguments presented by J. McDowell against the notion of highest common factor. I will defend that each one of those arguments, independently from each other, enables us to get to an externalist position.

KEYWORDS: externalism about the mental, private language, following a rule, two dogmas of empiricism, the third dogma of empiricism, highest common factor.

Introduction

Structure:

- Presentation of the paper's structure and its objectives. My central goal is to analyze some arguments that motivate externalism, which

are neither classical references within the externalist literature, nor are based on thought experiments. My parallel aim is to locate some of the externalist arguments present in Wittgenstein's, Davidson's and McDowell's positions. These philosophers are generally classified as being externalist, but it is not obvious which and how their arguments could favor such a position. Quine's argument is studied in a more modest manner and seen as somehow independent from Quine's overall position, although there are those (e.g., Davidson) who identify Quine as being a proper externalist. These issues are developed in four sections.

- The working hypothesis of this paper is the following: there are alternative ways of defending externalism, which don't need to be based on thought experiments. Such alternative paths are able to reveal more explicitly than a though experiment some of the philosophical questions involved in adopting an externalist position.

1. The Private Language Argument [P.L.A.] and the Rule-Following Argument [R.F.A.]

Most of the first section was published in English in *Reduction and Elimination: Proceedings of the 31th International Wittgenstein Symposium*. Kirchberg am Wechsel (Austria): Austrian Ludwig Wittgenstein Society, 2008, pp. 26-28. I've found it appropriate to transcript here entirely such a section instead of dealing only with its main topics.

I. Since Kripke has defended that "the real 'private language argument' [P.L.A.] is to be found in the sections preceding §243" (Kripke, 1982: 3) of *Philosophical Investigations* (PI), it has become an imperative –for those who want to enter the discussion– to figure out its relation to the rule-following argument (R.F.A).

In this paper, I will maintain that both arguments are in dialogue with each other, but not in the Kripkean sense. By doing

this, I will be able to offer a double externalist interpretation of them. On the one hand, P.L.A. –when considered as independent from R.F.A– will lead us to a negative formulation of the externalist thesis, through a *reductio ad absurdum* of the internalist conception of the mental. On the other hand, when both arguments are considered as concerning the same question, they will lead us to a positive defense of externalism.

I will take *externalism* to be the position that defends that mental contents and states are individuated with reference to external factors to one's skin and *internalism* to be the negation of this thesis.

II. A great part of the discussion about P.L.A. is centered in the case proposed by §258; a case where we are asked to imagine ourselves writing in a diary the occurrence of a certain private sensation. In this diary, we should write the sign 'S' every time we had that sensation. Wittgenstein warns us with respect to the traits of this exercise: "(…) The individual words of this language are to refer to what can only be known to the person speaking; to his immediate private sensations. So another person cannot understand the language" (PI §243).

The notion of private language criticized by Wittgenstein involves several questions; the question regarding completely private experiences (in the sense that no one could have access to them but their bearer), the question about the development of a language able to describe such experiences, and the question about the possibility of a language understood only by its creator. When Wittgenstein argues against the idea of a private language, he is arguing against such notions. Furthermore, he is arguing against a specific theory of language; one which supposes that an ostensive connection between a word and a sensation (or between a word and an object) is sufficient to establish meaning. §258 leads us to the final consequences of thinking in those terms:

> (…) A definition surely serves to establish the meaning of a sign.
> —Well, that is done precisely by the concentrating of my attention;

> for in this way I impress on myself the connection between the sign and the sensation. —But "I impress it on myself" can only mean: this process brings it about that I remember the connection *right* in the future. But in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about 'right'. (PI §258)

There are those who have interpreted such an argument as dealing with a skeptical problem about memory (Hacker, 1990: 108). Such an interpretation says that, although an ostensive definition can be made plausible, the problem is how to warrant the future connection between sensation S and its name. However, it seems that this kind of skeptical problem is not the core issue of Wittgenstein's argument (see Hacker, 1990: 180 and Gert, 1986: 429). In the case proposed by §258, the problem is not a question about knowing whether I apply the same word I am using now in the future, neither it is about how to remember the way I have used it in the past; more than that, the problem is that even in the current case we are not allowed to say that any meaning was established at all.

Another interpretation of P.L.A. is Kripke's one, where P.L.A. is only a particular case of R.F.A., an argument that according to him leads us to another skeptical paradox.

R.F.A. can be exemplified with the case proposed in §185. In such a case, a pupil is taught to write down the series of cardinal numbers of the form 0, n, 2n, 3n, etc., at an order of the form "+n". "So at the order '+ 1' he writes down the series of natural numbers" (PI §185). We are asked to suppose that the pupil has been tested up to 1000. Then, the pupil is asked to follow the series beyond 1000 with the order "+2". He writes 1000, 1004, 1008, 1012.

> We say to him: "Look what you've done!"--He doesn't understand. We say: "You were meant to add *two*: look how you began the series!"--He answers: "Yes, isn't it right? I thought that was how I was *meant* to do it."—Or suppose he pointed to the series and

> said: "But I went on in the same way."--It would now be no use to say: "But can't you see....?" —and repeat the old examples and explanations. (PI §185)

Kripke indicates that the core of R.F.A. is to demonstrate that "[a]dequate reflection on what it is for an expression to possess a meaning would betray (…) that that fact could not be constituted by any of *those*"; by any "available facts potentially relevant to fixing the meaning of a symbol in a given speaker's repertoire" (Boghossian, 1989b: 508). Under this interpretation, §185 proposes a skeptical paradox in similar terms to what seems to be suggested in the following aphorism:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here (…). (PI § 201)

Although this aphorism continues by saying that "It can be seen that there is a misunderstanding here from the mere fact that in the course of our argument we give one interpretation after another" (PI §201), Kripke insists on the skeptical scenario; a scenario that spreads to P.L.A.: nothing could fix the meaning of the sign 'S', like nothing could fix the meaning of the sign '+2' in the pupil's case. The solution found by Kripke to the supposed skeptical paradox is communitarism; if there is not such a thing as a semantic fact to determinate the difference between looking right and being right, deciding about this difference is something that belongs to the community.

McDowell (1984), however, who disagrees with Kripke's interpretation, offers us not just an important criticism to that interpretation, he also shows us another way of understanding Wittgenstein's position. McDowell stresses the conditions for the establishment of the skeptical paradox, insisting on the continuation of the §201:

> (…) What this shows is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases. (PI § 201)

McDowell holds that Kripke's paradox appears only if we keep considering meaning as an interpretation. The necessary step, therefore, would be to change the idea that understanding always involves offering an interpretation. That would be Wittgenstein's lesson. If R.F.A. does not concern the desperation of establishing the difference between right and wrong, the Kripkean conclusion is its aim either. If McDowell is right in his diagnosis, it is not the case that P.L.A. is just another instance where we can verify the skeptical paradox. In the case of the sign 'S', we are not allowed to say that we have established any meaning at all, but this is not the case with the sign '+2'. In a sense, both arguments are connected because they both dismiss the idea of meaning as being the univocal relation between a sign and an object, or between a sign and a mental image. However, they set apart in the sense that, the case of '+2' has a correction criterion, thought not established by a semantic fact, while in the case of 'S' it does not have it. In this sense, we could say that P.L.A. establishes a specific criticism to the idea of mental entities giving meaning to our language. Given this, we could return to P.L.A. and reformulate it in the following terms:

i. Possessing a correction criterion is a condition of possibility to a language;
ii. A private language lacks correction criteria;
iii. A private language is impossible. There is no such a thing as a private language because it is not a language.

The first premise refers to the idea that having a meaning is essentially a matter of possessing a correctness condition; a statement is meaningful if it can be true or false.

The second premise appears clearly at the end of §258. The attempt to point privately to a certain sensation, to a private one, leaves us without a correction criterion. The very sensation cannot

give me by itself such a criterion, as it seems to be presupposed by an ostensive definition between the sensation and the name I allegedly give to it. Wittgenstein rejects this image, not only here, but in most parts of *Investigations*. R.F.A. is an example of this rejection, but it appears also in the earlier aphorisms of PI, when Wittgenstein criticizes the Augustinian image of language. Given the two premises, the immediate conclusion of such an argument is that the "concept of a private language is one that cannot be defended, at best, and is incoherent, at worst" (Preti, 2002, 56).

P.L.A. has a deep externalist character. The notion of private language could indeed be elaborated in opposition to an externalist position: the components of such a language are not identified by external factors to an individual, but purely by internal ones. Because of that, to argue for the incoherence of such a notion makes room for reaching externalism through a *reductio ad absurdum*. And the conclusion is that it becomes unintelligible to talk, at the same time, about instances of language (it does not matter if we are talking about the world or about our subjective experiences) and about private correction criteria.

If by arguing in favor of P.L.A. we show the incoherence of internalism, this constitutes a motivation to reach externalism, though a negative one. However, it is also possible to find a positive motivation in Wittgenstein's arguments, by taking both P.L.A. and R.F.A. as working together. And this is possible if we think that, more than a criticism, they offer us an alternative option to think about meaning which does not need the idea of semantic facts.

Kripke seems to suggest that the Wittgensteinian argument leads us to communitarism. We could understand him as saying that premise (ii) is true because any correction criterion is to be established by a community. In this sense, one could find in Kripke's interpretation some semblance of externalist if we are able to retain the idea that individuating mental contents belong to the community and never to oneself privately. However, Kripke's position is much

stronger than that; the community is provided with full powers to the very establishment of meanings. While this position could sound as an externalism, it would also sound as the complete isolation of the community within itself. At this moment, "[o]ne would like to say: whatever is going to seem right to *us* is right. And that only means that here we can't talk about 'right'" (McDowell, 1984: 49, n. 12).

As I have tried to defend, not only Kripke's interpretation does not seem to be the most satisfactory one, but his solution also causes a discomfort towards which McDowell calls our attention. If in an internalist position we could be isolated from the community, now we could, all together, be isolated from the world. And this does not seem to be Wittgenstein's position, as Preti warns us:

> From the fact that our fellows in the community play a constitutive role in determining content it will not follow that content is *not* the "queer", inner mental process that Wittgenstein is concerned to deny. (…) Perhaps, that is, it is true that what determines meaning or content must be partly constituted by the minds of others – but it won't follow from this that the content in *other* minds in the community isn't determined by *their* inner mental processes. Merely being *other* is not enough to thwart the inner state conception of meaning, and it may be that Wittgenstein appreciated this. (Preti, 2002: 60)

There is, however, another way of making plausible the idea that correction criteria can only belong to the public sphere without the commitment to communitarism. And that is possible when we realize that the institution of meanings and their application are not distinctive activities. If the moments of application of meanings are so important in Wittgenstein's approach, this is so because they are not separated from the moments of institution of meanings. In this context, externalism is able to follow a positive path, and in that sense, different from the one reached with the accusation of incoherence of the notion of private language. Here the meanings

would be established with relation to external factors to one's skin, but also, with relation to external factors to any individual.

The positive character of Wittgenstein's argumentation is related to the dispute about the interpretation of his arguments; for example, regarding which notion of meaning Wittgenstein defends at all. A possible interpretation is the one developed by McDowell, who suggests that one learns to see the rules once one is part of a community; the community is the only instance that habilitates us to deal with the world linguistically (1984a). Prades (2006) supports the main lines of McDowell's interpretation, but points out that more need to be said in addition to the "introduction of a community for linguistic meaning as a consequence of the requirement that there must be a way of grasping meaning that should not be an interpretation" (Prades, 2006: 150). He indicates that such an additional element has to do with "the basic role that expressive behavior plays in the genealogy of content" (Prades, 2006: 149). Finkelstein highlights a related element, but before attributing to expression a central role, he emphasizes that the fundamental step taken by Wittgenstein is the dissolution of the idea that there is a gap between words and their meanings (or between intentions and actions, or even between expressions and mental states). Finkelstein defends that the crucial lesson to be learnt from the discussion about following a rule is that there is no such a gap. According to this interpretation, the paradox suggested by the first part of §201 is solved by the rejection of the impulse of seeing this gulf; a gulf between an order and its execution, for example (an image suggested in §431). In this sense, Finkelstein is able to retain the idea that the misunderstanding underlying the paradox of §201 is subsidiary of the idea that grasping a rule is to give an interpretation. According to Finkelstein (2008: 81), it makes sense to provide an interpretation when someone has misunderstood a sentence, for example, or when there is a real risk of misunderstanding. However, in the normal

case, we do catch the rule[1].

2. Arguments against the Two Dogmas (AcDD) – Willard O. v. Quine

Structure:

- Specification of the second section's objective: to outline the externalist consequences of Quine's argument against the two dogmas of empiricism; what is a more modest task than defending that Quine was a proper externalist, as Davidson (2003: 281) argues. Quine's controversial views about the mind (and his alleged nihilism about meaning) preclude us from identifying him automatically as an externalist. However, his arguments against the two dogmas –the analytic-synthetic distinction and reductionism– do seem to motivate an externalist account of the mind.

- Exposition of the distinction between analytic and synthetic sentences. A classical example of the first is "A bachelor is an unmarried man" and of the second "Snow is white". People who support such a distinction would probably say that, in the first case, the notion of unmarried man is already present in the meaning of 'bachelor', while in the second case, one would need further empirical information in order to decide about the truth value of the second sentence. The notion of analytic truth has at least two different senses: i. the predicate of an analytic sentence is contained in the subject; ii. an analytic sentence is true in virtue of its meaning, independently of how the subject matter is. Quine's argumentation rejects both senses.

1 I favor this interpretation and apply it in the sixth chapter of this dissertation in order to examine some questions related to self-knowledge. I also discuss the expressivist account with more attention in that chapter.

- Interpretation of Quine's criticism to such a distinction: first, there is no criterion to distinguish between analytic and synthetic sentences. Any attempt of defining the term 'analytic', for example, refers to other terms that also need clarification. Second, the very motivation to separate sentences whose meaning depend on the world (e.g., "Snow is white") and sentences which are allegedly independent of world information (e.g., "A bachelor is an unmarried man") is highly problematic. Such a motivation is informed by the idea that there are two moments in our linguistic life: one where language acts alone and another moment where it is the world which enters the scene.

- Externalist interpretation of the first part of the argument. The first dogma –the analytic-synthetic distinction– is consistent with the rejection of the externalist thesis. Such a dogma allows for the individuation of some mental contents (the ones that correspond to analytic sentences) independently from the world. By rejecting this dogma, one is allowed to defend that any sentence depends on the world, in the sense that it is necessary to refer to external factors to one's skin in order to identify one's mental contents. However, externalism emerges from the rejection of other three possibilities of understanding the dogmas:

i. I call 'hyper-internalism' the position which conceives the world in internalist terms, such as the theories based on sense data, from which the world arises as a logical construct. In this context, both analytic and synthetic sentences are individuated by internalist conditions. The inputs are, so to speak, in the subject's skin (or, alternatively, private sensations in her head). In this sense, holding the first dogma or rejecting it are both internalist enterprises. However, Quine's argumentation against the two dogmas provides us with elements to reject this conception of the world, especially by arguing against the dogma of reductionism.

ii. I call 'social internalism' the position that conceives

analytic sentences as already dependent on external factors –social ones– insofar as they are established in a public environment. In this context, rejecting the analytic-synthetic distinction would not represent a passage from internalism to externalism. However, appealing exclusively to the community seems to be rather a case of internalism, because analytic sentences also miss factual components. The division is established then between what is internal to a community of minds, such as conventions and definitions, and what is external to it, the world. In this sense, Quine's criticism of the first dogma applies both to the conception of meaning as a construct from private experiences and to the idea that analytic truths make explicit semantic conventions within a community, as if they were established without interference from the world.

iii. The third possibility it the one actually found in Quine according to Fodor and LePore (1992), which I call, following them, 'nihilism'. After rejecting the two dogmas, we reach confirmation holism and semantic nihilism: there are no semantic properties; there is no acceptable discourse about meaning. I argue that the dispute between the nihilist and the externalist result is not directly solved by elements found is Quine's argumentation against the two dogmas. They are found in the more radical rejection of the analytic-synthetic distinction offered by Davidson. I argue for this point at the end of this section.

- Exposition of the second dogma. The dogma of reductionism refers to the idea that the meaning of a statement can be reducible to some of its parts, the ultimate atoms (e.g., sensorial inputs); the elements that supposedly have a direct connection to the world (Quine, 1951: 85).

- Criticism to the second dogma. The second dogma is intimately connected with the first one, and in this sense, it can be prey of a similar criticism: we cannot separate nor make intelligible that a

part of our language is connected to the world and the rest is not. Quine dismisses the possibility of talking about ultimate atoms, by emphasizing that "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body" (Quine, 1951: 85).

- Externalist interpretation of the second part of the argument.
The reductionist thesis seems to be in principle compatible with the idea that some contents –at least the ultimate atoms– are individuated with reference to the world, since they are supposed to have direct contact with it. However, having a direct connection to the world doesn't imply an externalist commitment, since 'hyper-internalism' is an open possibility. Furthermore, sensorial inputs can be understood as being located *in* the subject's skin (not beyond it). However, even if sensorial inputs were to be individuated in a externalist manner, the problem is that reductionism depends on the previous dogma. It presupposes that the ultimate atoms have null contribution from the subject's conceptual system (or from the conventions of a community), depending exclusively on the non-conceptual content of the subject's experience. At the same time, reductionism works if there are circumstances where statements are "vacuously confirmed, *ipso facto*, come what may; and such statement[s] [are] analytic" (Quine, 1951: 85). Rejecting such a dogma motivates the idea that world and linguistic factors are both necessary to identify one's mental contents. The holism that emerges from the rejection of the two dogmas supports the idea that a sentence's truth value depends both on linguistic and on world factors. As I've mentioned before, our statements seem to be judged as a corporate body. Any change in any element of our belief system will have consequences for other items. In this sense, such a proposal favors externalism insofar as the world factors (understood as a generic term) are required in order to identify any and all elements of a belief system.

- Conclusions.

Rejecting the two dogmas allows us to deny both hyper-internalism and social internalism. However, after rejecting the two dogmas, we are left with two positive solutions: nihilism and externalism. Rejecting the two dogmas implies that we cannot separate the world's contribution from language's contribution, but there are also two interpretations of such a conclusion: 1. to understand that those contributions cannot be separated although they remain intrinsically different, and 2. to maintain that those contributions cannot be separated because they are not different in kind. The first interpretation seems to be the base both for nihilism and two-factor externalism (as a variation of the one defined in the first chapter), in the sense that our systems of beliefs as a whole are populated with internal and external elements to a subject (or to a community). However, this seems to be a version of the analytic-synthetic distinction, but in holistic terms. The second interpretation of the relation between the world's and language's contributions emerges from the rejection of the holist version of the analytic-synthetic distinction. That we cannot demarcate where language's contribution begins and where it ends means that elements without a propositional form cannot interfere in our system of beliefs. Something as sense data is unintelligible because it is unintelligible to conceive that world factors could affect our system of beliefs as a whole, without having a propositional form. These are certainly not Quine's conclusions, but Davidson's ones, who also rejects the third dogma of empiricism. However, since this conclusion is reached through the rejection of the analytic-synthetic distinction in its strong version (the holistic one), Quine's position seems to be unstable. Rejecting the two dogmas seems to tend to the rejection of the third one if someone considers the ultimate consequences of rejecting the analytic-synthetic distinction. In this sense, the two-factor externalism which emerges from the traditional reading of the two dogmas' rejection also seems to tend to the global externalism

found in Davidson.

3. Arguments against the Third Dogma (AcTD) – Donald Davidson

Structure:

- Specification of the third section's objective: to outline the externalist consequences of Davidson's argument against the third dogma of empiricism. This argument doesn't exhaust Davidson's externalist elements, although it is central to such a position. Davidson's externalism is properly discussed in the third chapter of this dissertation.

- Exposition of the dualism between conceptual scheme and empirical content. That notion is what Davidson calls 'third dogma of empiricism'. The main idea involved in such a dualism is that our cognitive and linguistic activities organize the brute material received from the world through our senses. This dualism has actually two referents: the separation between concepts and naked sensations, on the one hand, and the separation between world vision and the world itself, on the other hand.

- Exposition of Davidson's criticism to the dualism, which seem to have the following argumentative structure:

1. Conceptual schemes can be said to be different if there is some degree of untranslatability between them.
2. There are no cases, complete or partial, of untranslatability.

CONCLUSION: We cannot talk about different conceptual schemes.

*

1'. The distinction between conceptual scheme and empirical content can be sustained if there is a distinction among

schemes or if it exists one unique scheme that contrasts with the world.

2'. There are no differences among conceptual schemes (in the sense of the above conclusion)

3'. There is not a unique scheme that contrasts with the world.

CONCLUSION': there is no distinction between conceptual scheme and empirical content.

The first part of the argument is sustained by the idea that we cannot conceive someone as a linguistic being and, at the same time, judge such a person as having an unintelligible language; as having a completely different world view from our own. Premise (1) makes explicit the definition of 'different conceptual schemes'. Premise (2) is defended through the idea that recognizing someone as having a world view requires her sharing of most of our beliefs (Davidson, 1974: 197). In a radical situation, such as the case proposed by Quine where a translator intends to translate the language of a completely isolate community, one is required to identify which sounds are samples of linguistic behavior and which are not. In this sense, the interpretation of meanings and of mentality occur at the same time (Davidson, 1974: 195). Once one tries to understand the meaning of 'Gavagai', one presumes to be facing intentional behavior. According to Davidson, Quine's story about radical translation is incomplete insofar as it misses this basic point about how translation involves interpretation, and therefore, attribution of mentality. Quine's rejection of the two dogmas is also incomplete under Davidson's view, because rejecting such dogmas is still compatible with the possibility of a community changing completely the parameters of its belief system, till the point of having untranslatable schemes. The rejection of the first two dogmas, if not accompanied by a rejection of the third, makes room for a holistic form of the analytic / synthetic distinction, being now within language, thought or science as a whole where a separation between the world's contribution and the

subject's (or the community's) contribution can be made. This result makes room only for two-factor externalism.

The second part of the argument includes Davidson's overall view about interpretation. Premise (1') makes explicit the two possibilities of sustaining the scheme / content distinction. Premise (2') is justified by the first part of the argument. Premise (3') is defended by Davidson through the principle of charity: interpretation presupposes that interpreter and interpretee share a great part of beliefs, and the interpreter is required to attribute a great degree of correctness to the person interpreted. Furthermore, Davidson argues, most of such beliefs must in fact be true. More details about the defense of premises (2) and (3') are found in chapter 3 of this dissertation.

-Externalist interpretation of Davidson's rejection of the third dogma.

If the scheme / content dualism (e.g., between the world and world views) is rejected, the intrinsic connection between world and mind is reached. And this can be seen as the general externalist consequence present in this argument. Our world view is not isolated from other world views, neither from the world itself. A language or a linguistic community is the very proof that world content and knowledge are constitutive to them. That is to say, it is only possible to talk about mind once we conceive it as constituted by knowledge from the very beginning. I favored such a position in Chapter 1.

4. Rejection of the Highest Commom Factor (MFC) – John McDowell

Structure:

- Specification of the fourth section's objective: to outline the externalist consequences of McDowell's argument against the

notion of highest common factor. This argument doesn't exhaust the externalist elements of McDowell's philosophy. Neither does it necessarily entail his disjunctivist theory of experience.

- Exposition of the notion of highest common factor: the common element between a real experience and a mere hallucination. For example, in an experience of seeing something red, there would be three elements: the red object in the world, the mental image of such a red object in the mind, and the mind that accesses such an image. A hallucination of a red object would allegedly involve the same mental image of the real experience. And discerning between the two situations is impossible for the subject of such experiences. The notion of highest common factor rejected by McDowell concerns three main issues: the conception of an intermediate between mind and world; the idea of a mind accessing its internal objects; and an internalist conception about attribution of knowledge.

- Transposition of the notion of highest common factor to the Twin Earth thought experiment.
The assumed common factor between Oscar's experiences about water and Twin Oscar's experiences about twin water has the same characteristics of the highest common factor between a real experience and a hallucination: it acts as an intermediary between mind and world, it is conceived as a sort of mental image internally accessible to the subject, and it precludes both individuals from knowing exactly in which situations they are. As I've exposed in the first chapter of this dissertation, the part of Oscar's mind that differs from Twin Oscar's corresponds to broad states and the part that remains the same is taken to be narrow; that is, individuated independently from external factors.

- McDowell's criticism of the highest common factor.
Conceiving the mind as populated by internal objects –the

intermediate elements between mind and world– is useless and inaccurate. According to McDowell, it supposes a conception of experience in terms of satisfaction of criteria, which is highly unsatisfactory. For example, experiencing water and experiencing twin water would be satisfied by the same criteria: the appearance of water. We could reason about the attribution of mental states also in terms of satisfaction of criteria. In this case, we would have the following situation: a pain behavior and a simulation of it are indiscernible because they satisfy the same criteria: someone's appearing to having pain. "But since 'criteria' are defeasible, it is tempting to suppose that to experience the satisfaction of 'criteria' for a claim may not be true. That yields this thesis: knowing that someone else is in some 'inner' state can be constituted by being in a position in which, for all one knows, the person may not be in that 'inner' state. And that seems straightforwardly incoherent" (McDowell, 1982: 371).

Besides McDowell's insistence on how problematic is this conception of experience, the rejection of the highest common factor also involves the dissolution of a sort of internalist account on the conditions of knowledge's attribution. One doesn't need to know one's own situation in order to be attributed with knowledge or with ignorance. That is, one doesn't need to discern between water and twin water in order to be attributed and, in fact, to think in terms of twin water and water.

- Modest externalist interpretation of McDowell's rejection of the highest common factor.
By dissolving the idea of the mind as populated by intermediaries between it and the world, one is able to dismiss the idea of narrow contents.

- Robust externalist interpretation of McDowell's rejection of the highest common factor.

In order to replace the conception of experience in terms of satisfaction of criteria, McDowell proposes a disjunctive notion of it. With this move, he is able to defend that a real experience and a hallucination are different from the very beginning. In twin earth cases, for example, Oscar experiences water and Twin Oscar experiences twin water. There is no similarity between the two situations. And this sort of direct contact between mind and world is what I've called the 'robust externalist interpretation' of McDowell's rejection of the highest common factor. Under this view, one cannot individuate one's own mental states without referring to the world. The alleged mental images appealed to in order to identify one's experiences have been dissolved. At least in the veridical cases, the world is certainly present in one's mind.

Conclusion

All the arguments studied in this chapter make explicit, the incoherence of internalism, by following different paths and involving a variety of philosophical questions. I've interpreted some of them as offering also a positive outlook regarding how to pursue externalism.

3 Davidson's Externalisms

Cristina Borgoni and Herivelto Souza.

ABSTRACT
Donald Davidson has deeply contributed to what is nowadays called Externalism. However, the exact formulation of his externalism is not obvious since his externalist commitments are spread along many of his papers. The aim of this work is to explore the details of his externalism. We will point out that Davidson clearly defends that the mind is not self-contained. Nonetheless, this idea acquires at least two different senses under his view: on the one hand, mental states and contents must be individuated in part by factors external to one's skin because the former were caused by the latter; and on the other hand, mental states and contents must be individuated in part by external factors because the mind is constituted by knowledge. We will point out that the apparently harmonious relation between those two levels of explanation turn out to be conflictive at a certain point within the very Davidsonian program.

KEYWORDS: Donald Davidson, externalism, radical interpretation, mental causation, triangulation.

RESUMEN
Donald Davidson ha tenido un papel extremamente importante en lo que hoy se llama Externismo. Sin embargo, la formulación exacta de su externismo no es obvia porque sus compromisos están dispersos a lo largo

de muchos de sus artículos. El objetivo de este trabajo es explorar los detalles de su externismo. Indicaremos que Davidson sin duda defiende que la mente no está auto-contenida. No obstante, tal idea tiene por lo menos dos sentidos distintos en su trabajo: por un lado, estados y contenidos mentales deben ser individuados en parte con respecto a factores externos a la piel de uno porque fueron causados por ellos; y por otro lado, estados y contenidos mentales deben ser individuados por factores externos porque la mente está constituida por conocimiento. Indicaremos que la relación entre estos dos niveles explicativos, aparentemente harmoniosa, se vuelve conflictiva dentro del mismo programa davidsoniano.

PALAVRAS CLAVE: Donald Davidson, externismo, interpretación radical, causación mental, triangulación.

Introduction

Donald Davidson has deeply contributed to what is nowadays called 'externalism', a philosophical position about the mind, which has its most refereed roots in Hilary Putnam's (1975) and Tyler Burge's (1979) papers. In several of his works, Davidson has stressed that the mind is not self-contained, that it is not independent of the world. He suggests indeed that the mind turns out to be unintelligible within an internalist framework. However, the exact location of his externalism is not obvious since the theses which compose it are spread along many papers.

The aim of this work is to pursue such a location through indicating the externalist theses which Davidson is committed to. We will point out that Davidson clearly defends the externality of the mind. Nevertheless, such an externality gains at least two different senses or explanations under his view, which could serve to characterize two different ways of motivating externalism. We will be especially interested in characterizing the relation between both

kinds of "externalisms".

We begin by raising a sort of conflict between Davidson's thought experiment of Swampman and his theses about language and radical interpretation. Such an experiment asks us to imagine a creature, Swampman, which is interpreted by others and seems to interpret them though it misses a mind. Davidson's arguments involved in his theory of interpretation and in his denial of the third dogma of empiricism preclude him to conceive such a creature. Departing from this conflict, we intend to trace a route along some of Davidsonian theories and theses in order to delineate his externalism.

First, we will discuss the Davidsonian idea of triangulation understood as the requirement of causal connections between the individual, her community and the world in order for the mind to emerge. We will then introduce his view about linguistic practices as being the very proof of the idea of triangulation. Once we are interpreters and are interpreted, and as long as we ascribe beliefs and meanings mutually to each other, the fundamental connections between ourselves, the community and the world must have been already established, for there is no sense in talking about mind and about mental contents in the absence of any of the vertexes.

Our next step will be to examine his denial of the third dogma of empiricism, the distinction between conceptual scheme and content. With this movement, Davidson is able to conclude that the mind cannot be detached from the world, in the sense that we are knowers. Such a position stresses once more the compulsory role of the community as well the compulsory involvement of the world. The reasons that sustain such a denial, however, respond to his theses about radical interpretation, which will be discussed in the subsequent section.

In the sixth section, we will propose two senses by which we can understand Davidson's externalism: on the one hand, mental states and contents must be individuated in part by factors external

to one's skin because the latter were caused by the former; on the other hand, mental states and contents must be individuated in part by external factors to one's skin because the mind is constituted by knowledge. Those two senses will correspond to two levels of explication about the externality of the mind.

In the conclusive section we will raise a question about what could prevent us from accepting Swampman as an open possibility within Davidson's externalism. We will offer an overall view about Davidson's theory of anomalous monism, which seems to be in agreement with the scenario where Swampman is created. We will try to defend that nothing intrinsic to such a discourse poses any problem to the viability of a creature such as Swampman. And so we will be back to the tension initially indicated by the paper.

1. Let's suppose during one of Donald Davidson's stays in England, he decides to give himself a break and go for a cruise along Thames Estuary, located in a swampland area. While he is waiting for his boat it starts to rain and, suddenly, lightning strikes a dead three besides him. Entirely by coincidence, Davidson's body is reduced to the tree elements and the tree turns into Davidson's physical replica. This replica, Swampman, moves exactly as Davidson used to do. He gets into the boat, appears to enjoy the trip and go back to London, where he was giving a series of philosophical interviews. Swampman seems to recognize Davidson's friends, appears to return their questions in English and seems to manage all philosophical theses Davidson used to sustain. No one could tell the difference.

> But, there *is* a difference. My replica [Davidson says] can't recognize my friends; it can't *recognize* anything, since it never cognized anything in the first place. It can't know my friends' names (though of course it seems to) (…). It can't mean what I do by the word 'house', for example, since the sound 'house' Swampman makes was not learned in a context that would give it the right meaning – or any meaning at all. Indeed, I don't see how my replica can be said to

> mean anything by the sounds it makes, nor to have any thoughts. (Davidson, 1987: 19)

In other words, Swampman has no mind. Davidson's mental states have a causal history while Swampman hasn't any, since it appeared only a few minutes ago. According to such an experiment, despite Swampman being interpretable and seeming to be able to interpret others, *it* is a *thing, it* has no mind.

But then, one could recall the theses involved in the theory of interpretation and the denial of the third dogma of empiricism advanced by Davidson. In those contexts, Davidson argues against the very idea of alternative conceptual schemes. He argues that once we recognize a conceptual scheme, that is, when we recognize someone as a linguistic being, we are already sharing most of our beliefs with such a person. He sustains that "given the underlying methodology of interpretation, we could not be in a position to judge that others had concepts or beliefs radically different from our own" (Davidson, 1974: 197).

Such a methodology of interpretation involves, on the one hand, the interdependence between belief and meaning because, for Davidson, the attribution of beliefs and the interpretation of sentences occur at once. On the other hand, it requires the assumption of a general agreement between interpreter's and speaker's beliefs in addition to the assumption of a great deal of correctness about the speaker's beliefs (Davidson, 1974: 195-6).

When we interpret someone, Davidson sustains, we need to keep some things stable. In an extreme case of radical interpretation, the linguist wouldn't be able to give even a first step towards translation if she doubted simultaneously the meaning of the utterance *Gavagai*, the native's accuracy when he says *Gavagai,* and the nature of the sound as constituting a linguistic behavior[1]. According to Davidson, even an ordinary conversation requires keeping something fixed. In

1 We will discuss Quine's example in more detail in section 5.

order to interpret a sound as intentional, we must suppose that there is someone who intends to say something; that we are dealing with a minded individual. In such a situation, to doubt the mindedness of the creature seems to be blocked.

The Swampman case, however, seems to open the way to turn alternative conceptual schemes intelligible. It is an example of a completely empty scheme with which any other subject would share no beliefs. In that scenario anyone who interprets Swampman is wrong, since there is nobody to be interpreted. This certainly highlights an important tension within Davidson's framework because in many of his writings, Davidson holds that the mind exists inasmuch as it is attributed, and it is attributed inasmuch as it in fact exists.

This may be one of the reasons why Davidson has regretted, in so many passages, the use of such "science fictions" in order to delineate his philosophical positions. And such an idea is our starting point: if the case for externalism can be made with Swampman, it can be even better made without[2].

2. It seems that if we want to understand Davidson's externalism, it would be more appropriated not to depart from the Swampman case, since it represents –to say the least– a case of total failure of translatability, what is precisely denied by Davidson in several works. Nevertheless, we could retain some of the theses that surround Davidson's thought experiment in order to reach his externalist commitments, such as the idea that mental states and contents must be individuated appealing to external factors, because the former

2 We are following Davidson's own observations about his thought experiments such as "I also agree (…) that the argument that summons up an Omniscient Interpreter does not advance my case. As with Swampman, I regret these sorties into science fiction and what a number of critics have taken to be theology. If the case can be made with an omniscient interpreter, it can be made without and better" (Davidson, 1999a: 192).

are caused by the latter. What is more, the external character of the mental involves all the history of such connections instead of isolated events, as supposed by Putnam. According to Davidson, such a change in his approach supports the denial of mental's division between broad and narrow contents, and consequently allows the spread of externalism to the mind as a whole.

The classical Twin Earth experiment offers a situation where Oscar and Twin Oscar use the same word with different meaning in spite of their inability to access the "true" meaning of their terms. From this case, Putnam concludes that psychological states don't determine the extension of the terms. Davidson's criticism towards him indicates that Putnam could have lead externalism much further than he actually did. Davidson highlights that the difference between Oscar and Twin Oscar's causal history precludes us from considering them as being in the same psychological states.

Putnam's route towards externalism begins with his diagnosis of a bad philosophical tradition that has insisted in sustaining simultaneously the following assumptions:

> (I) That knowing the meaning of a term is just a matter of being in certain psychological state (...)
>
> (II) That the meaning of a term (in the sense of 'intension') determines its extension (in the sense that sameness of intension entails sameness of extension). (Putnam, 1975: 219)

Putnam's externalism emerges from the denial of (I), which charges him with the burden of keeping the correspondent psychological states as narrow ones. Davidson insists that the sort of paradox pointed out by Putnam takes place only when we depart from an internalist view about psychological states. The broadening of externalism conducted by Davidson covers not only meanings, but psychological states in general as well.

When Davidson sustains that the causal history of our terms is what should count for the individuation and the determination

of mental contents, he is able to maintain that Oscar and Twin Oscar's minds should be distinct in all aspects, including those states that supposedly provide material for self-knowledge. According to Davidson, "it doesn't follow, simply from the fact that meanings are identified in part by relations to objects outside the head, that meanings aren't in the head" (Davidson, 1987: 31) in the sense that the subject wouldn't know his own thoughts if they were individuated in an externalist manner. Davidson states that "to suppose this would be as bad as to argue that because my being sunburned presupposes the existence of the sun, my sunburn isn't a condition of my skin" (Davidson, 1987: 31). Two burned skins could be visually indistinguishable, but if one of them was caused by the sun and the other not, we should take into account such external factors in order to individuate both injuries. One is a case of sunburn while the other is not.

3. The teacher is responding to two things: the external situation and the responses of the learner. The learner is responding to two things: the external situation and the responses of the teacher. All these relations are causal. Thus the essential triangle is formed which makes communication about shared objects and events possible. But it is also this triangle that determines the content of the learner's words and thoughts when these become complex enough to deserve the term. (Davidson, 1990: 203)

The ideas exposed in (2) acquire a better application when understood under the perspective of Davidson's notion of triangulation. There can be no mind without any of the vertexes that form the triangle: the individual, the community and the physical world. Since these three factors constitute necessary conditions for the emergence of thought, and since they contribute to determining its contents, the causal relations between those vertexes must be taken into account for the individuation of mental states and events.

However, the triangulation story is neither so brief, nor reduced

to the talk about causal relations in isolation. Davidson's view about interpretation is reflected here once more. He sustains that our linguistic life is the very proof of the idea of triangulation. Once we interpret and are interpreted, that is, as long as we ascribe beliefs and meanings mutually to each other, the fundamental connections between ourselves, the community and the world must have already been established.

Davidson insists on the fact that the distinctive aspect of his concept of triangulation, as a tool to indicate necessary (although not sufficient) conditions for thought and talk, is the introduction of the community pole. Since triangulation can be taken also as a sort of sketch of language acquisition[3], the second individual that stands for community does the important job of helping to identify the relevant cause of an utterance in a given situation. Quine had already shown that it was not sufficient for the learner to hear a sentence in the presence of an object in order to grasp its meaning: it is essential that the learner sees that the teacher also sees the object (Quine, 1969b: 28). The problem, Davidson argues, is that Quine's "epistemology remains resolutely individualistic" (Davidson, 2001: 10), for he has always insisted on the role of sensory stimulation (the triggering of nerve endings) as the only clue to the meaning of an observational sentence. But who knows something about the patterns of stimulation of another person? In contrast, Davidson noticed early that the option for the distal cause of the stimulus was the only that could serve that purpose. That's why Davidson defends that

> Our triangular model thus makes a step toward dealing with another troublesome feature of Burge's perceptual externalism, the

3 As in Quine, for whom the radical translation situation was analogous to the coming of a child into the language practices of a community, the Davidsonian triangulation can be considered as a description of what is in play in radical interpretation as well as a structure of the elements involved in the process of language learning.

> indeterminate nature of the contents of perceptual beliefs. That difficulty arose because there seemed to be no way to decide the location of the objects and features of the world that constitute the subject matter of perceptual beliefs; Burge told us only that the content was given by the 'usual' or 'normal' cause. But this did not help choose between proximal and distal stimuli, or anything in between, in the causal chain. By introducing a second perceiver, it is possible to locate the relevant cause: it is the cause common to both creatures, the cause that prompts their distinctive responses. (Davidson, 2001: 8-9)

In order to manage that, Davidson seems to suggest that individuals have some kind of natural ability[4] to associate another creature's responses to features of the shared world that possibly have caused such responses. "We are built to discriminate objects, to keep track of them, expect them to emerge from their hole or from behind trees, and in some cases to feed or eat us", says Davidson (1999b: 731). But then, once one gets mastery on concept use, and so becomes able to propositional thought, Davidson suggests that "perception is propositional: when we look or feel or hear we believe. What we are caused by our senses to believe is often true, which in the simplest cases it could not fail to be, since the content of our simplest beliefs is necessarily fixed by the history of past perceiving" (Davidson, 1999b: 732).

So, this makes clear that, according to Davidson, an individual wouldn't have thoughts without having language, which has a social basis in the sense that "without one creature to observe another, the triangulation that locates the relevant objects in a public space could not take place" (Davidson, 1990: 202). Such a social life, which is needed for having a mind, also requires a good deal of actual knowledge, not only of features of the objectively common

4 In fact, one way to conceive this is supposing that such abilities have emerged as result of processes like natural selection. What is important, anyway, is to identify factors that are necessary to make intelligible how thought could ever come out, and this kind of ability was certainly in play there.

environment, but of other people's minds as well. The external aspect of the mind is explicated, on the one hand, by the requirement of relations between oneself, her community and the world, and on the other hand, by the mind's being constituted by knowledge. It seems that if causal relations between the three vertexes of the triangle are enabling conditions for the mind, the presence of knowledge is another important requirement in Davidson's account.

4. That the mind cannot be detached from the world, in the sense that we are knowers, is a conclusion reached by Davidson when he rejects the third dogma of empiricism, the distinction between conceptual scheme and content.

Davidson accuses such a dualism to be committed to the idea that our cognitive and linguistic abilities are organizing activities of a brute material offered by the world through experience. "Conceptual schemes, we are told, are ways of organizing experience; they are systems of categories that give form to the data of sensation; they are points of view from which individuals, cultures, or periods survey the passing scene" (Davidson, 1974: 183). According to him, this image is all one needs to conceive different conceptual schemes, different systems of concepts that could be untranslatable to one another, even though sharing the same objective world. In such a scenario, "Reality itself is relative to a scheme: what counts as real in one system may not in another" (Davidson, 1974: 183).

When Davidson refuses the dualism between scheme and content, he undermines the possibility of isolating conceptual form from empirical content, as well as distinguishing between supposed radically different conceptual schemes. His reasons respond basically to the unintelligibility of the idea of relative or alternative conceptual schemes:

> For what is the common reference point, or system of coordinates, to which each scheme is relative? Without a good answer to this question, the claim that each of us in some sense inhabits his own

> world loses its intelligibility. […] The meaninglessness of the idea of a conceptual scheme forever beyond our grasp is due not to our inability to understand such a scheme, nor to our other human limitations; it is due simply to what we mean by a system of concepts. (Davidson, 1988b: 39-40)

So, a system of concepts cannot be systematically separated from the world in which it was formed. The Davidsonian attack on the third dogma, then, has a double movement: "one of them criticizes the separation of concepts and naked sensations. The second rejects the divorce between world-views or schemes and the universe" (Pinedo, 2004: 271).

According to Davidson, when Quine (1951) denounced the two first dogmas, he banned the image of several worlds seeing, heard or described from the same point of view, a conception of world-views that could dispense with the contribution of the world itself. With this move, Quine provides a kind of warranted contact to the world by concluding that the total belief system of a given community gets affected by the world through experience, at least at its periphery. Nevertheless, Davidson argues that the rejection of the two dogmas doesn't prevent us from imagining a single world seen and described from different points of view that would also be untranslatable one into another[5].

The rejection of the third dogma is specifically directed towards such a kind of relativism, since even abandoning the two dogmas, as suggested by Quine, we still do not rule out the idea that a community could change completely the standards of its system of beliefs to the point of creating untranslatable conceptual schemes.

Renouncing the third dogma, as suggested by Davidson, also provides a sort of warranted contact with the world, but dispenses

5 That is, there can be always a translation between conceptual schemes, in Quine's view, but there is a constitutive indetermination in any translation, which keeps the specter of systematic mistake, or insuperable incommensurability, haunting us. That's why we can say they might be untranslatable.

with any reference to experience; at least as empiricists conceive it, as having a primary role. Davidson defends that anything that deserves the name of world-view is inseparable from the world that is viewed, and this is enough for one not to require any empiricist explanation of such a contact. The mind is conceivable only as being constituted in the presence of the world.

> Given the dogma of a dualism of scheme and reality, we get conceptual relativity, and truth relative to a scheme. Without the dogma, this kind of relativity goes by the board. Of course truth of sentences remains relative to language, but that is as objective as can be. In giving up the dualism of scheme and world, we do not give up the world, but re-establish unmediated touch with familiar objects whose antics make our sentences and opinions true or false. (Davidson, 1974: 198)

The denial of the dualism between concepts and naked sensations means that only within a conceptual environment –within a linguistic community– one could individuate mental contents. According to Davidson, once the third dogma is rejected, there are no reasons to maintain empiricism. The well known Davidsonian slogan "only beliefs justify beliefs" (Davidson, 1983) stresses the idea that any possibility of content's individuation is carried out by pertaining to a community, and that precludes any appeal to a privileged moment when the world reveals itself barely, as in the Quinean tribunal of experience.

The second important result of rejecting the third dogma is the abandonment of the divorce between world-views and the world itself. Such a conclusion is made apparent when Davidson states that once there is something that suggests the existence of a conceptual scheme, such as a language or a linguistic community, world content is already there. Whatever we identify as being our world vision cannot be completely isolated from other world-views, nor from the world itself.

In this sense, the rejection of the third dogma turns compulsory

the role of a community as well as the participation of the world. There are not several possibilities of organizing the supposed brute material offered by the world, all untranslatable between each other. The world itself must be present in our actual understanding activities, even if we don't have absolute warrants about which part of our belief systems is more accurate than others.

5. The compulsory participation of both the world and the community sustaining the talk about mental contents is a conclusion afforded by Davidson's account on radical interpretation.

Quine (1960) asks us to imagine a case of a community completely isolated from any other on earth and, in addition, a linguist who goes to such a community in order to elaborate a translation manual. The linguist observes numerous situations, for example, one where a rabbit runs in front of them and the natives say 'Gavagai'. The linguist takes note of this verbal behavior and translates it as 'rabbit'.

Quine stresses that such a translation hypothesis will be tested in similar future circumstances in order to corroborate the linkage between the stimulus and the verbal behavior. In addition, given the plurality of stimuli that could have provoked such an emission of sound, the hypothesis at hand will be put under the approval of someone from the community. In that case, the linguist should be able to perceive the native's assent or dissent in front of the question "*Gavagai?".* Given the complexity of this whole situation, Quine concludes that the indetermination of translation should be the case. Mistakes and failures could be managed or diminished, but the method of translation should be considered in fact as inconclusive if our aim was the search for synonymies.

Nevertheless, Davidson considers that the situation proposed by Quine has many other features that should receive more of our attention. According to Davidson, the very activity of the linguist in such a community –taken at first as having a completely

different conceptual scheme– would require the satisfaction of some important conditions, such as the linguist's ability for perceiving what could count as being verbal behavior. Davidson emphasizes that to perceive something as a conceptual scheme isn't given to us at all. Perceiving some behavior as a contentful linguistic one puts us automatically in a comprehensive position of such a conceptual system, in the sense that it cannot be completely isolated from our own:

> Radical interpretation establishes that something meaningful cannot be understood in isolation from other meaningful things, but rather globally: when we make sense of someone's speech or rationalize her behaviour we need to assume a shared world which is inconceivable independently of a shared intentional net. Davidson dedicates to this idea one of his most subtle arguments: his rejection of the third dogma of empiricism. (Pinedo, 2006: 11)

In the absence of such a shared net of beliefs, one couldn't even recognize verbal behaviors as meaningful assertions in the so considered isolated community. One couldn't recognize such a community as a linguistic one. That is the key of the Davidsonian insistence in the interdependency between meanings and beliefs.

The attribution of beliefs and the interpretation of someone's words occur simultaneously, as aspects of the same practice. That is why Davidson maintains that someone's speech cannot be interpreted unless a good part of what the speaker's beliefs are about is known by the interpreter. The elaboration of a translating manual doesn't require only a good match between native words and ours. It is an activity not only of translating other's language into ours, it involves the description of other's attitudes as well (Davidson, 1974: 186). And to describe someone's attitudes is an activity of ascription of mental states, such as beliefs and desires.

Davidson sustains that the basis of our interpretation activities is the sharing of beliefs and, in addition, it requires from us the attribution of correctness to the person we are interpreting. The

extreme situation of radical interpretation stresses the myriad of mistakes the linguist is subject to. She could be wrong about her choice of translation and the native could be wrong in using a term of his own language. However, Davidson insists, interpretation is only possible if we don't take into account all the possibilities of mistake involved in any situation. Because of this, Davidson defends that we need to ascribe a great deal of success to our interpretee[6], a condition known as the principle of charity:

> Since charity is not an option, but a condition of having a workable theory, it is meaningless to suggest that we might fall into massive error by endorsing it. (...) Charity is forced on us; whether we like it or not, if we want to understand others, we must count them right in most matters. (Davidson, 1974: 197)

> The method is not designed to eliminate disagreement, nor can it; its purpose is to make meaningful disagreement possible, and this depends entirely on a foundation – some foundation – in agreement. The agreement may take the form of widespread sharing of sentences held true by speakers of 'the same language', or agreement in the large mediated by a theory of truth contrived by an interpreter for speakers of another language. (Davidson, 1974: 196–7)

We have seen that sharing beliefs constitutes an enabling condition for interpretation. We have also seen that such an idea gives place to the rejection of a solipsistic isolation between different conceptual schemes. The so called 'relative world views' cannot be enclosed within themselves. Following this same line of reasoning, we could also say that Davidson provides us with means to reject the idea of a speaking individual isolated inside her own world. Given that the conditions for radical interpretation apply to foreign situations as well as to domestic ones, we are able to state the impossibility of one's own isolation from every other individual. There is no sense in talking about private worlds.

6 This does not mean we have to ascribe success in all situations, but that the ascription of error is intelligible only against the background of massive success.

Besides sharing beliefs and attributing truth to other's beliefs in order to interpret them, Davidson defends that radical interpretation requires that a great deal of those beliefs must be indeed true. This fact turns knowledge as another enabling condition for interpretation, an important one which by itself prevents us from being isolated from the world. Davidson defends that:

> Until a base line has been established by communication with someone else, there is no point in saying one's own thoughts or words have a propositional content. If this is so, then it is clear that knowledge of another mind is essential to all thought and all knowledge. Knowledge of another mind is possible, however, only if one has knowledge of the world, for the triangulation which is essential to thought requires that those in communication recognize that they occupy positions in a shared world. So knowledge of other minds and knowledge of the world are mutually dependent: neither is possible without the other. […] Knowledge of the propositional contents of our minds is not possible without the other forms of knowledge since there is no propositional thought without communication. It is also the case that we are not in a position to attribute thoughts to others unless we know what we think since attributing thoughts to others is a matter of matching the verbal and other behaviour of others to our own propositions or meaningful sentences. Knowledge of our minds and knowledge of the minds of others are thus mutually dependent. (Davidson, 1991: 213)

6. At this point, we already have important clues to understand Davidson's externalism. On the one hand, Davidson justifies the individuation of the mental with reference to external factors because our mental contents and states are caused in some sense by those factors. I am in relation to my community and to the world in a way that those relations provide externality to my mind. On the other hand, the individuation of the mental with reference to external factors is justified because the mind is constituted by knowledge, a

mental state that by itself could not lack an external aspect[7].

This double aspect of Davidson's position suggests two levels of explications about the externality of the mind[8], which could be characterized respectively by the following conditions:

EC1 (explanatory condition 1): mental states and mental contents must be individuated in part by external factors to one's skin because the former were caused by the latter.

EC2 (explanatory condition 2): mental states and mental contents must be individuated in part by external factors to one's skin because the mind is constituted by knowledge.

Those explanatory conditions seem to refer to matters so distinct from each other that they deserve to be treated separately. On one level, the mind has an external character because we've been in causal relations with the world and with the community, while on the other level the mind is not self-contained because it is necessarily composed by knowledge. Both conditions constitute Davidsonian externalism, but it is useful to recognize that EC1 provides an extrinsic relation between mind and world while EC2 gives space to an intrinsic one[9].

In the previous sections we have seen that our interpretation

7 Williamson (2000) indicates that part of the resistance to externalism lies in the insistence of defining knowledge. According to Williamson, knowledge should not be considered as the result of the articulation between something internal (belief) and something external (truth) through justification. The very notions of beliefs and justification depend for their intelligibility on the notion of knowledge. Although Davidson defends that we are knowers, it is not clear whether Davidson would accept Williamson's approach to knowledge.

8 We are deliberately not considering the division between "social externalism" and "perceptual externalism", suggested by Davidson (2001), as an important trait of his externalism, since it's clear that Davidson insists on the necessity of the community and of the world. We are inclined to think that such a division is superficial and disguises the relevant differences between externalisms.

9 Davidson recognizes that externalism is an alternative to subjectivism when it "makes the connection between thought and the world intrinsic rather than extrinsic – a connection not inferred, constructed, or discovered, but there from the start" (Davidson, 2001: 2).

activities only stand in the presence of knowledge. Davidson's rejection of the third dogma involves an important aspect of our communicative activities that is the intrinsic connection between world visions and the world itself. There is no way to give sense to the idea of an alternative conceptual scheme because once we recognize someone as having a conceptual scheme (or, to avoid the jargon, as being a conceptual creature) we are inevitably attributing to her, and so sharing with her, meanings and beliefs at once. There is no sense in considering her as an intentional being, and still regard her as an unintelligible speaker in principle[10]. One cannot understand my beliefs and simultaneously believe that all of them are false. Besides sharing beliefs, Davidson concludes that most of such beliefs should be indeed true. Knowledge is so identified as a necessary condition for interpretation.

Davidson (1991) stresses that knowledge of the world is necessary, but so are knowledge of other minds and self-knowledge. The interdependence between these three kinds of knowledge is another important condition defended by Davidson. But one may ask at this point: if the mental is conceived as necessarily composed by knowledge, why insist in extra conditions to justify the external character of the mind? If our interpretation activities are evidence of our having knowledge, it seems that such a fact would be enough to explain the external character of the mind. Once we take such a step there is no way back to conceive the mind as detached from the world.

Davidson clearly defends both levels of explications and more than that, he seems to indicate that EC2 is subsidiary of EC1. In the Davidsonian picture, that the mind is necessarily constituted by a good portion of knowledge seems to respond to the idea that

10 She may turn out to be unintelligible for other reasons, which include, for example, some kinds of disease, but she cannot be unintelligible because her mental contents and meanings are intangible. See Davidson (1973).

we are embedded in causal relations[11]: the objectivity of the mind emerges from the triangulation conceived as composed by causal relations. One way of interpreting Davidson's insistence on such an aspect is seeing it as a grounding of the epistemological and semantic features of his position on a metaphysics that is coherent with his commitment to physical monism. In this sense, causal relations can be read as stressing the fact that all the events are physical ones: that's why the triangular relations are ultimately causal. And even if Davidson has always made clear that causal relations do not bear semantic content, it may not be so simple to render triangulation intelligible, as presenting the explanatory role it has, once one insists that such an ontology of purely causal physical events makes some features of mind and meaning quite mysterious[12].

7. But, what does prevent us from conceiving Swampman as an open possibility within Davidson's externalism? We have refused to take such an experiment as the representative case of Davidson's externalism because amongst other things Davidson himself has claimed that it is not a good example of it. However, it was also indicated that more than being a non-representative case, Swampman expresses an important internal tension within Davidson's framework.

The Davidsonian solution to the mind-body problem –his anomalous monism– is a position that respects the following principles:

i. Principle of Causal Interaction: "at least some mental events interact causally with physical events" (Davidson,

11 It seems that both explanatory lines could be sustained separately. Williamson (2000) offers us an example of such a possibility when he considers knowledge as being a mental state which is prior conceptual and metaphysically to the one of belief.

12 Consider, for instance, what Chalmers (1995) coined as "the hard problem of consciousness".

1970: 208).

ii. Principle of the Nomological Character of Causality: "where there is causality, there must be a law: events related as cause and effect fall under strict deterministic laws" (Davidson, 1970: 208).

iii. Anomalism of the mental: "there are no strict deterministic laws on the basis of which mental events can be predicted and explained" (Davidson, 1970: 208).

Although such principles could be considered as controversial when held together, Davidson defends that there is a way of maintaining all of them simultaneously by embracing anomalous monism. Amongst the theses involved by such a position there is one regarding individuation, according to which an event is individuated with respect to its causes and its effects (Davidson, 1969); and also one establishing that it is possible to talk about causes and effects using a physical vocabulary as well a mental one (Davidson, 1963).

Davidson maintains monism about events by stating that there are not two classes of events, one physical and the other mental. Instead, there is just one kind of events that could have two descriptions, one physical and another mental.[13] He stresses that

13 It is important to emphasize that Davidson's work largely assures autonomy for the mental vocabulary. This is noticeable in several papers. In his (1973b), Davidson asks us to consider the existence of Art, a robot built as a perfect physical replica of a human being. He says "If we want to decide whether Art has psychological properties, we must stop thinking of him as a machine we have built and start judging him as we would a man" (Davidson, 1973b: 251; see Davidson 1974b for related issues). Davidson's framework, however, is different from a framework such as the one developed by Dennett (1979, 1987), which characterizes mental vocabulary in terms of the application of an intentional strategy; a strategy chosen among other available explicative strategies, to predict and to explain an object's or a system's behavior. While neither Davidson nor Dennett put the emphasis on ontological considerations to argue for their positions, Davidson, but not Dennett, takes it as a precondition for the attribution of mental states that the interpretee herself is a user of intentional vocabulary. Furthermore, it can be argued that the ineliminatibility of mental state attribution is Davidson's way to resist the temptation of thinking of ourselves as primarily subject to prediction

every mental event has a physical description but not the other way round. Davidson sustains a Humean notion of causality which allows him to retain the discourse about laws, but at the level of events, not of its descriptions. Davidson's central idea is that causal relations are established between events, independently of their descriptions, although a nomological law could only appear under a physical description. Such a position entails that a mental event could be described physically, and then could be part of causal relations even if the idea of psycho-physical laws is rejected (Davidson, 1967). This way, Davidson maintains the anomalous character of the mental while he is able to sustain the identity between physical and mental states.

The Swampman experiment is not disconnected from this part of Davidson's philosophy. In the same context where he proposes Swampman, he indicates that one of the reasons why people consider it so difficult to conceive psychological states as external ones lies in the fact that nobody has countenanced an approach like anomalous monism. According to Davidson, once it is possible to talk about psychological states within the sphere of causality, the supposed problem seems to be solved[14]. Considering that Swampman has no causal history, neither physical nor mental, and considering that any mental state that pertains to the real Davidson is part of a causal net that the replica lacks, their minds must be completely different (assuming it makes sense to speak of Swampman's mind).

At this moment, it seems clear that nothing intrinsic to this discourse poses any problem to the viability of a creature such as Swampman being taken in our linguistic practices. And we are back to the tension initially indicated by the text. Our impression is that

and control (Ramberg, 2000:366-67). Maybe the introduction of the personal stance (1976) approximates Dennett and Davidson, but this would be a question for further investigation.

14 See Davidson (1971) and (1973c) for a discussion of the relation between causality and agency.

if the second explanatory level of externalism –that which refers to the constitution of mind by knowledge– is subsidiary of the first one, EC2 arrives too late to disallow the conception of a creature such as Swampman, mainly because in interpreting its utterances, part of what one has to concede is that those words were learned in the presence of causal relations. And if such a creature lacks those relations, and still presents all that is required to count as a minded being, it will certainly be attributed with propositional thought. There's no way of checking embeddedness in causal relations in past history of learned language. And the point is that checking embeddedness is certainly not part of the history.

EC2 is the externalist explanatory condition which maintains that "mental states and contents must be individuated by external factors because the mind is constituted by knowledge". It was suggested that such a level of explanation is able to make room for an intrinsic relation between mind and world, since knowledge is a sort of mental state that could not be conceived as detached itself from the world. Nevertheless, Davidson seems to suggest that EC2 is dependent of another level of explanation, EC1.

EC1 outlines the idea that mental states and contents are caused by external factors and, because of that, the former must be individuated by the latter. This provides an external character to the mind due an extrinsic relation between mind and world. EC2 is related to EC1 in the sense that EC1 is more fundamental than EC2. That the mind has knowledge within its constituents is a result of causal relations between an individual, her community and the world. Only in a second stage interpretation seems to come into the scene. In that sense, when Swampman is finally taken to be an unintelligible creature, he is already there; he has been interpreted and has interacted with our fellows, even lacking a mind. Needless to say that skepticism concerning other minds does not get blocked in this explanatory strategy, what would require turning it upside down, and giving priority to EC2. In fact, that seems to be Davidson's

position after having regretted the thought experiment: from the fact that mind is constituted by knowledge, it is possible to infer that language was learned in the presence of causal relations. This sort of movement may assure the rejection of positions such as the ones defending that "the views generally called 'externalist' do not form a particularly interestingly interconnected family of theses" (Hahn, 2003: 29). Locating Davidson's commitments to externalism, as we pursued to do, had in view precisely presenting a picture of why Davidson could so clearly emphasize that "what I think is certain is that holism, externalism, and the normative feature of the mental stand or fall together" (Davidson, 1995: 122).

Bibliographical References

Burge, T. 1979, "Individualism and the Mental", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 21-83.

Chalmers, D. J. 1995, "Facing Up to the Problem of Consciousness", *Journal of Consciousness Studies* 2 (3), pp. 200-19.

Davidson, D. 1963, "Actions, Reasons, and Causes", in Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 3-19.

Davidson, D. 1967, "Causal Relations", in Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 149-162.

Davidson, D. 1969, "The Individuation of Events", in Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, 1980, pp. 163-80.

Davidson, D. 1970, "Mental Events", in Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 207-27.

Davidson, D. 1971, "Agency", in D. Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 43-61.

Davidson, D. 1973, "Radical Interpretation", in Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 125-39.

Davidson, D. 1973b, "The Material Mind", in D. Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 245-59.

Davidson, D. 1973c, "Freedom to Act", in D. Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 63-81.

Davidson, D. 1974, "On the Very Idea of a Conceptual Scheme", in Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 183-98.

Davidson, D. 1974b, "Psychology as Philosophy", in D. Davidson 1980, *Essays on Actions and Events*, Clarendon Press, Oxford, pp. 229-39.

Davidson, D. 1983, "A Coherence Theory of truth and Knowledge", in Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 137-53.

Davidson, D. 1987, "Knowing One's Own Mind", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 15-38.

Davidson, D. 1988b, "The Myth of the Subjective", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 39-52.

Davidson, D. 1990, "Epistemology Externalized", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 193-204.

Davidson, D. 1991, "Three Varieties of Knowledge", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 205-20.

Davidson, D. 1995, "Can There Be a Science of Rationality?", in D. Davidson 2004, *Problems of Rationality*, Clarendon Press, Oxford, pp. 117-34.

Davidson, D. 1999a, "Reply to A. C. Genova: The Very Idea of Massive Truth", in L. E. Hahn (ed.) 1999, *The Philosophy of Donald Davidson*, Chicago, Open Court, pp. 192-94.

Davidson, D. 1999b, "Reply to Dagfinn Følllesdal", in L. E. Hahn (ed.) 1999, *The Philosophy of Donald Davidson*, Chicago, Open Court, pp. 729-32.

Davidson, D. 2001, "Externalisms", in P. Kotatko, P. Pagin & G. Segal (eds.) 2001, *Interpreting Davidson*, Stanford, CSLI Publications, pp. 1-16.

Dennett, D. C. 1976, "Conditions of Personhood", in D. Dennett 1981, *Brainstorms: Philosophical Essays on Mind and Psychology,* Cambridge, Mass., MIT press., pp. 267-285.

Dennett, D. C. 1979, "True Believers", in D. Dennett 1987, pp.13-35.

Dennett, D. C. 1987, *The Intentional Stance*, Cambridge, Mass., MIT Press.

Hahn, M. 2003, "When Swampmen Get Arthritis: 'Externalism' in Burge and Davidson", in M. Hahn & B. Ramberg (eds.) 2003, *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, Cambridge, MIT Press, pp. 29-58.

Pinedo, M. 2004, "The Anomalous Character of Experience", in J. C. Marek & M. E. Reicher (ed.) 2004, *Experience and Analysis* (Proceedings of the 27th International Wittgenstein Symposium), Kirchberg am Wechsel (Austria): Austrian Ludwig Wittgenstein Society, pp. 269-71.

Pinedo, M. 2006, "Anomalous Monism: Oscillating between Dogmas", *Synthese* 148, pp. 79-97.

Putnam, H. 1975, "The Meaning of 'Meaning", in H. Putnam 1975, *Mind, Language and Reality,* Philosophical Papers, vol. 2, Cambridge, Cambridge University Press, pp. 215-71.

Quine, W.v.O. 1951, "Two Dogmas of Empiricism", *The Philosophical Review* 69, pp. 20-43.

Quine, W.v.O. 1960, *Word and Object*, Cambridge, Mass., MIT Press.

Quine, W.v.O. 1969b, *Ontological Relativity and Other Essays,* New York, Columbia University Press.

Ramberg, B. 2000, "Post-ontological Philosophy of Mind: Rorty versus Davidson", in R. Brandom (ed.) 2000, *Rorty and his Critics*, Oxford, Blackwell Publishers, pp. 351-377.

Williamson, T. 2000, *Knowledge and its Limits*, Oxford, Oxford University Press.

4 When Externalism and Privileged Self-knowledge are Compatible and When They are not[1]

Cristina Borgoni. To be published in *Episteme NS*, vol. 29, num. 1, 2009

ABSTRACT

This paper focuses on the incompatibilist debate between externalism and privileged self-knowledge, such as it appears in the literature under two privileged contexts of discussion: the slow-switching cases and the *reductio ad absurdum* arguments. My aim is to defend a compatibilist position although recognizing some exceptions to it. I will defend, on the one hand, that the incompatibilism reached by slow-switching cases is sustained only in case we maintain a specific but problematical view about self-knowledge. On the other hand, the incompatibilism reached by *reductio ad absurdum* arguments is only sustained if we maintain a narrow conception of externalism.

KEYWORDS: externalism, self-knowledge, incompatibilism, slow-switching cases, *reductio* of compatibilism.

---

1 I would like to thank Manuel de Pinedo and Sarah Sawyer for comments and suggestions on all aspects of this paper. I am also grateful to Jesús Palomo Muñoz and César Marini. This work has been partially supported by the research project HUM2007-63797/FISO (MEC). A previous version of this argument will be published in Portuguese in Silva Filho, W. (ed.) *Mente, Linguagem e Mundo: O significado do anti-individualismo e autoconhecimento*. São Paulo: Alameda.

RESUMEN
Este trabajo se dedica al debate incompatibilista entre externismo y autoconocimiento privilegiado, tal como aparece en la literatura bajo dos contextos específicos de discusión: los casos de transferencia lenta (*slow-switching cases*) y los argumentos de tipo *reductio ad absurdum*. Mi objetivo es defender una posición compatibilista que a la vez reconozca algunas excepciones a ella. Defenderé, por un lado, que los casos de transferencia lenta logran dar bases a un incompatibilismo solamente si mantenemos una visión específica pero problemática acerca del autoconocimiento. Por otro lado, los argumentos de tipo *reductio ad absurdum* logran dar bases a un incompatibilismo solamente si mantenemos una concepción estrecha del externismo.

PALAVRAS CLAVE: externismo, autoconocimiento, incompatibilismo, casos de transferencia lenta, *reductio* del compatibilismo.

Introduction

As Ludlow (1998: 1) suggests, externalism is in a sense "the denial of the traditional Cartesian view that holds that the contents of our thoughts are what they are independently of the surrounding world". What philosophers normally take to be Cartesianism[2] is not only committed to internalism about mental content, but mainly to a view on self-knowledge according to which such a realm plays a fundamental role both in epistemology and in metaphysics. Under

2 After Burge defended that Cartesianism was committed to internalism, or better, that "Individualism as a theory of mind derives from Descartes" (Burge, 1986: 192), he portrayed himself as a defender of the anti-individualism spirit in Descartes's works (2003/2006). This seems to be an open question, but as far as a kind of established caricature called 'Cartesianism' exists –a position committed both to individualism about mental contents and to self-knowledge entirely acquired by direct and non-empirical means– I will refer to such a theoretical position when talking about Cartesianism.

this approach, self-knowledge is conceived as a kind of knowledge entirely acquired by privileged means, that is, in a way that dispenses with any empirical investigation and with any inferential process.

Following this reasoning, since externalism is incompatible with Cartesianism, one should expect externalism to be incompatible also with the possibility of privileged self-knowledge[3]: one thesis or the other obtains, but not both. However, as we all know, Cartesianism is neither the only nor the best available account of self-knowledge, even if we don't forfeit its special trait such as its acquisition by us in a direct and non-empirical manner.

This paper is dedicated to the incompatibilist debate between externalism and privileged self-knowledge, such as it appears in the literature under two favored contexts of discussion: the slow-switching cases and the *reductio ad absurdum* arguments[4]. My aim is to defend a compatibilist position although recognizing some exceptions to it. I will defend, on the one hand, that the incompatibilism reached by slow-switching cases is sustained only in case we maintain a specific but problematical view about self-knowledge. On the other hand, the incompatibilism reached by *reductio ad absurdum* arguments is only sustained if we maintain a narrow conception of externalism.

In the first part, I shall discuss some incompatibilist arguments based on the thought experiment of slow-switching and their respective compatibilist answers. I will also discuss the role such thought experiments play in the general context of the discussion, indicating that a compatibilist answer could be designed without the consideration of such cases. At this point, I will compare Tyler Burge's and Donald Davidson's compatibilism.

---

3 The term 'privileged self-knowledge' will be used here in reference to the direct and non-empirical way by which we acquire at least part of our self-knowledge. It is important to notice that both aspects – directness and non-empiricism – will be required to characterize such a specific knowledge.

4 Respectively referred by Davies (2000: 391) as 'the achievement problem' and 'the consequence problem'.

In the second part, I will treat the reduction arguments. I will defend that the externalist premise over which those arguments are constructed is misleading in relation to what an externalist is committed to.

1. The Slow-Switching Cases

The thought experiment that establishes the first discussion context is exposed by Burge in 1988, in a paper where he defends a compatibilist position. It is the "thought experiment of slow-switching cases", where a subject –let's say Oscar– is stealthily shifted back and forth between actual Earth and Twin Earth, several times, remaining unaware of those shifts. Oscar acquires the appropriate concepts to each situation, such as water and twin water. If Oscar is told about such switches and asked to identify when they took place, he will not be able to answer.

Boghossian's comments (1989a) on slow-switching cases have given rise to two different incompatibilist arguments: one that emphasizes the discrimination of mental contents from their relevant alternatives and another one which emphasizes the question about memory.

1.1. Discrimination of mental contents and relevant alternatives

The first of those incompatibilist arguments can be restructured as follows:

> (P1) To know that P by introspection, S must be able to introspectively discriminate P from all relevant alternatives of P.
>
> (P2) S cannot introspectively discriminate water thoughts from twin water thoughts.
>
> (P3) If the Switching Case is actual, then twin water thoughts are relevant of water thoughts.
>
> (C1) S doesn't know that P by introspection. (Warfield, 1992: 218)

This argument stresses the consequences of being unable to distinguish between actual and twin situations. The underlying intuition is that in order to have knowledge of any content, one should be able to distinguish it from the relevant possibilities. The thoughts Oscar has on Twin Earth establish relevant alternatives to the thoughts he has on Earth. However, Oscar is unable to distinguish between them only by introspection.

On the basis of such an argument there is an important distinction between merely logical alternatives to one's thoughts and relevant ones. In the standard externalists' scenarios, such as the one proposed by Putnam (1975), Twin Oscar's thoughts about twin water represent only logical alternatives to Oscar's thoughts about water. In those cases, to require Oscar to discriminate his water thoughts from twin water thoughts would establish an implausible condition to knowledge, such as discriminating one thought from every single alternative possibility to it. However, in slow-switching cases, insofar as the subject of the switches seems to have both concepts[5], to discriminate between them seems to be more acceptable.

As Boghossian points out, "the ordinary concept of knowledge appears to call for no more than the exclusion of 'relevant' alternative hypotheses [...] and mere logical possibility does not confer such relevance" (Boghossian, 1989a: 158). In order to know that I have €2,25 in my pocket, I do not need to have checked that there is no forgery money in the vicinity, nor do I need to be able to tell the difference between a genuine euro and every imaginable forgery to that (Boghossian, 1989a: 158). I just have to count the coins. But, if I had 20p together with my Euros, I should be able not to count them. In this case, differentiating Euros from Pounds seems to matter to my final knowledge.

Following Boghossian's argument, Oscar would not have

5 There are important nuances in the interpretation of the case that the subject possesses two concepts. This question will arise later on in this paper.

knowledge by introspection of his own thoughts about water because he is unable to distinguish them, also by introspection, from his other thoughts about twin water. It seems that in order to distinguish them –and so, to know them– Oscar would have to engage in an empirical search, which will be favorable to incompatibilism.

However, this first version of Boghossian's comments has no effect at all on Burge's compatibilist strategy developed in 1988. His position consists in showing that there is a class of self-knowledge, named as the 'basic' one, which would resist such proofs. We would not need to differentiate the items of this class from their relevant alternatives because this group has the characteristic of being self-verifying. In this case, (P1) would be false[6].

Despite the inability to discriminate between twin periods from home ones, Burge will argue, the subject of the experiment is still able to have privileged self-knowledge, at least in reference to the so called '*cogito*-like judgments': a range of second-order thoughts which are "self-verifying" because of their self-referential form, such as "I think that I am thinking that water is wet".

The appeal to the *cogito*-like judgments guarantees that, at least in a specific range of self-knowledge, we can find not only the externalist aspect of mental contents, but also a special way to acquire them. Burge's position is, in fact, somewhat stronger than that, because he takes *cogito*-like judgments to be the paradigmatic instances of self-knowledge. That is why he identifies them as the 'basic self-knowledge'. Although this class of knowledge plays such an important role in Burge's position, it is crucial to notice that, according to him, not all self-ascription of beliefs are self-referential or self-verified. He maintains that a variety of self-knowledge cases extend out of what he has called the basic one[7] (Sawyer, 2002:

6 Furthermore, (P1) is false in Burge's account because, according to him, discriminating between relevant alternatives plays a more decisive role in empirical judgments than it does in self-knowledge.

7 Sawyer (2002) points out that Burge's thesis doesn't constitute a general theory

112). However, this special class has some epistemic peculiarities. According to Burge:

> The source of our strong epistemic right, our justification, in our basic self-knowledge is not that we know a lot about each thought we know we have. It is not that we can explicate its nature and its enabling conditions. It is that we are in the position of thinking those thoughts in the second-order, self-verifying way. Justification lies not in the having of supplemental background knowledge, but in the character and function of the self-evaluating judgments. (Burge, 1988: 660)

Burge insists that this specific group of second-order thoughts are self-referential, and hence self-verified, partly because the first-order thought, externalistically individuated, is somehow embedded in the second-order thoughts. This idea has constituted the most accepted answer to this first formulation of incompatibilist worries. According to Davies, in order to sustain compatibilism, several positions have based his answer on the fact that:

> [W]hen I think that I am thinking that water is wet, I deploy in thought the very same concepts of water and of being wet that are involved in my thinking that water is wet. So an externalist dependence thesis that is true for my first-order thinking that water is wet will be no less true for my second-order thinking that I am thinking that water is wet. Because the content of my second-order thought embeds the content of my first-order thought, my second-order thinking shares the dependence on the environment that is characteristic of my first-order thinking. (Davies, 2000: 391)

This point will be discussed again in the following sections, but one thing is important to retain. The appeal to this kind of condition to the second-order thoughts does not exactly mean following Burge in his answer to the incompatibilist problem. His answer is quite stronger insofar as it is sustained by the conception of a specific self-

---

of authoritative self-knowledge. Nevertheless this is not necessary for providing a good compatibilist answer.

knowledge class, the basic one[8].

As a result of this, one could insist that the restricted group to which Burge refers doesn't satisfy incompatibilist worries. It would be necessary to talk about self-knowledge in general. More than that, basic self-knowledge could not even resemble what we would like to take as the representative group of self-knowledge. So, just as a theoretical device, let's exclude Burge's strategy for a while in order to understand a little more about the core of this sort of incompatibilist argument.

Ludlow (1995a) not only considers Boghossian's argument (as restructured by Warfield) to be a cogent one, but also proposes a stronger reading of it. He adds the following premise to the argument above:

"(P4) Switching cases, in general, are prevalent" (Ludlow, 1995a: 227).

Ludlow claims that because "we routinely move between social groups and institutions, and in many cases shifts in the content of our thoughts will not be detected by us" (Ludlow, 1995a: 228), we are subject to situations very similar to those proposed by slow-switching cases. He maintains that departing from what he identifies as 'social externalism' – "namely that content is socially determined and that the relevant social groups may be highly localized" (Ludlow, 1995a: 229) – premise (P4) is entirely plausible. In doing so, Ludlow considers that Boghossian's argument can be better defended.

Ludlow exemplifies his position with the English word 'chicory', which designates two different but seemingly similar vegetables in England and United States. And he imagines a British traveler

8 We could indeed see Burge's proposal as composed by two elements: the reference to basic self-knowledge (which offers an evidence that there is self-knowledge in an externalist scenario) and the reference to the "embedding condition" (which explains that the concept employed in the second-order thought is the same as the one employed in a first-order thought that is individuated by externalist conditions).

who constantly goes from one country to the other. In addition, the traveler remains long enough in the United States and consequently acquires the mental content related to that environment, in such a way that the traveler would have his thoughts shifted each time he enters each country, remaining however unaware of this.

Although Ludlow (1995a) presents an interesting line of argument when he brings thought experiments to our daily lives, it seems that his thesis about the prevalence of those cases is misleading. For it seems that, in order to really imagine a situation where a person remains unaware of the double aspect of a word while she uses it constantly, we would have to imagine her completely isolated of any social contact. Let's think of another word, for example, 'chips', which in Britain means strips of potato fried in deep fat (which Americans call 'French fries'), while in the United States, the land of poker, it means casino tokens. It is very unlikely that a person –Carol, for example– coming from the United States, might use such a word without originating an initially conflictive situation that could be easily solved at a certain moment. Whenever she hears something like "these chips are delicious with vinegar", she would inevitably learn the second use of the same word.

In this way, the fact that there are some daily situations where we may find similarities to slow-switching cases doesn't mean that we are subject to them most of the time. What is more, the requirement of remaining long enough in the other environment in order to acquire the mental contents related to such a place is not a mere question of passage of time. It refers exactly to the fact that while the subject was there, there were interactions between her and the objects of such an environment, as well as between her and the people from that place. The problem seems to be that Ludlow assumes that the subject could be maintained inert to such interactions. Externalism is, however, exactly the opposite of such an idea. Those interactions constitute the very subject and her new experiences.

Considering the way we learn new words, it seems that such

learning includes indeed knowledge about the different contexts where they are used. That is why, several times, we prepare ourselves not to use some words or expressions when we are going to other countries, other cities or even to other social contexts. We usually know that they have other meanings elsewhere. Furthermore, we normally realize when we are entering a different social group or engaging in a different language game. Externalism doesn't require our ignorance about different environments.

Let's go back to Boghossian's argument in its original version. It seems that an important criticism is the one developed by Warfield (1992 and 1997).

Warfield criticizes such an argument stating that "Boghossian's shows at most that those individuals who are being slow switched fail to know the contents of some of their thoughts" (Warfield, 1997b: 232). That is, conceding some extent of soundness to Boghossian's argument, it doesn't go any further than showing that "externalism is *consistent with* a lack of self-knowledge; it does not show that externalism *implies* a lack of self-knowledge" (Warfield, 1997b: 232).

According to Warfield:

> To show that these doctrines are incompatible one needs to show that every possible world in which externalism is true is a world in which individuals do not have privileged self-knowledge. Boghossian shows at most that some possible worlds are worlds in which externalism is true and individuals lack privileged self-knowledge and Ludlow [1995a] shows at most that one world, the actual world, is a world in which externalism is true and (*some*) individuals lack privileged self-knowledge. (Warfield, 1997b: 233)[9] [my italics]

9 Ludlow understands this condition as follows: "Warfield's insistence that I show privileged self-knowledge to be false in every possible world in which externalism is true completely inverts the argumentative burden here." (Ludlow, 1997: 236). I would insist that the burden of the proof is in fact with the incompatibilist because, on the one hand, slow-switching cases are not but abnormal situations, and by other side, one cannot reach incompatibilism from those cases unless one defends a specific approach to self-knowledge.

Even in the stronger version defended by Ludlow – which I have considered as misleading – incompatibilism would not hold for similar reasons. Ludlow's conclusion was that most of us are most of the time under slow-switching cases and because of that, most of us fail to know our own thoughts if we take externalism to be true. But the conclusion required in order to reach incompatibilism would be that, considering externalism as true, all of us fail to know by privileged means every single thought we have.

At this moment, the impression is that the argument doesn't fulfill Boghossian's or Ludlow's expectations. Showing some cases where externalism is taken as true while individuals lack self-knowledge doesn't prove anything else than a compatibilism between externalism and a lack of self-knowledge; a conclusion not only acceptable, but quite accurate. It seems that there is enough data favoring the idea that we lack such an easy knowledge about all our thoughts. Failures of self-knowledge such as self-deception and akrasia seem to be merely the extreme cases that corroborate such an idea. On the one hand, privileged self-knowledge seems to be plainly true, but on the other hand it seems that we need to recognize that some range of self-knowledge is acquired by other manners than the privileged one.

In that sense, there seems to exist one situation where Boghossian's incompatibilism would work: if one insisted that there could not be a case of self-knowledge which was not potentially knowable to the subject in a direct and non-empirical manner. In slow-switching cases, self-knowledge about one's water thought may fail at a certain moment.

The fact that Oscar is unable to discriminate between his water thoughts and his twin water thoughts is not likely to affect *cogito*-like judgments. If Oscar states "I'm thinking that I think that fish live in the water", he is probably right about what he is thinking. However, it can affect self-knowledge if we consider another sort of examples.

Let's suppose Oscar states "I believe that I understand that

fish breathe in water because I also believe that fish' gills are able to extract oxygen from water". It seems that this case fails to be a good piece of self-knowledge in slow-switching cases, since in Twin-Earth water is not composed by H_2O. In this case, Oscar is mistaken about his own understandings[10]. His thoughts about his own understanding could be corrected by an expert in Twin Earths, but this would certainly require the acquisition of further information about one's environment. And this fact, in Boghossian's argument, leads to incompatibilism. But again, the unacceptable point of the incompatibilist argument seems to be that Oscar is required to always be able to know his own thoughts in a direct and non-empirical way. However, appealing to this condition in order to deal with privileged self-knowledge is neither required nor acceptable. Lots of times we are aware of our thinkings in an indirect way, because someone has called our attention to some aspect of our behavior or because we ourselves have engaged in some kind of self-analysis.

Therefore, the thought experiment in question – where the maintenance of the externalism is the supposed reason to the failure of one stance of privileged self-knowledge – could be used to sustain incompatibilism only in case we were assuming a very specific conception of self-knowledge: wherever it is part of this realm it must be entirely knowable a priori and directly. It seems that nowadays we have a lot of data favoring the denial of this conception[11].

10 That fish breathe in water and that fish breathe in twin water are both true. In this sense, Oscar's belief that fish breathe in water will be true, whatever concept he employs, water or twin water. That fish extract oxygen from water is true in Earth but false in Twin Earth. Oscar would have a true belief in case he employs the concept water but false in case he employs the concept twin water. Because of this, his reasoning about his own understanding is mislead. Once Oscar has both concepts, he cannot make the link between his belief that fish breathe in water and his belief that fish extract oxygen from water without knowing which one he is using.

11 This kind of compatibilist answer doesn't constitute an approach to self-knowledge, neither this is the aim of this paper. However, it suggests an important condition to an approach that wants to maintain externalism at the same time: it

1.2 Memory

The second line of argument favoring incompatibilism attempts to show that once slow-switching takes place, and externalism is considered to be true, there is no way of making sense of the memory of one's own thoughts. Several authors have found this argument also in Boghossian's comments (1989a) and can be restructured as follows:

> [P1'] If S does not forget anything, then whatever S knows at time t1, S knows at time t2.
>
> [P2'] In the cases at hand S does not forget anything.
>
> [P3'] S does not know that p at time t2.
>
> [C4'] So S does not know that p at time t1. (Burge, 1998a: 356)[12]

---

should be able to accommodate both methods of acquisition of self-knowledge.

12 Burge considers this the only interpretation of Boghossian's argument. He says: "Much of the literature on this subject deals with problems that arise from the assumption that we need to *identify* the content of our thoughts in such a way as to be able to rule our relevant alternatives to what the content might be. Boghossian, unlike many of those who write on this subject, seems to recognize that this assumption is not acceptable on my view. One's relation to one's content, when one is non-empirically self-attributing in the reflexive, that-clause way is not analogous to a perceptual, identification relation to which alternatives would be relevant. In present tense self-attributions of the relevant kind, alternatives are irrelevant. Boghossian's strategy is to consider cases of memory and argue that these cases reflect badly on my view about the present tense cases" (Burge, 1998: 355). However, if one reads through Boghossian's comments it seems that his argument does offer a double interpretation, especially because in the first formulation, Boghossian dedicates a good space to differentiate cases where relevant alternatives matter and where the problem is only about logical possibilities. As far as both formulations have received equal importance, I am considering both as valid.
Brueckner (1997) does not only agree with Burge's reading, but he also thinks that Boghossian's argument is directed to Burge's account of basic self-knowledge. Nevertheless, Brueckner will conclude that "no Boghossian-style argument succeeds in refuting Burge's account of basic self-knowledge" and adds to it: "the covariation strategies are untouched as well" (Brueckner, 1997: 330). I have considered that Boghossian's argument is much more general than applied just to basic self-knowledge.

Let's remember Oscar, the subject that undertakes the switches. Let's suppose that just after one set of twin-earthian concepts has been displaced by a set of earthian ones, someone were to ask Oscar whether he had been recently thinking thoughts involving an arthritis-like concept distinct from arthritis. He would presumably say 'no' (Boghossian, 1989a: 160). But the fact is that, according to externalism, Oscar does entertain thoughts which involve twin concepts. The question that arises here is about how to explain this sort of "poor" ability to know past thinking. Since this does not seem to correspond to a bad capacity of remembering them properly, Boghossian suggests that in those cases, Oscar in fact never knew them.

Boghossian claims that although Burge is able to say that at t1 Oscar knows what he is thinking at that moment, he must accept that at t2 Oscar would fail to know what he was thinking at t1, exactly because the self-verifying character of basic self-knowledge applies only to current thoughts. Boghossian understands that:

> By Burge's criteria (…) [S] counts as having direct and authoritative knowledge at t1 of what he is thinking at that time. But it is quite clear that tomorrow he won't know what he thought at t1. No self-verifying judgment concerning his thought at t1 will be available to him then. (Boghossian, 1989a: 171)

Again, it seems that Oscar would have to discover features of his environments in order to know what he himself thought in t2, exactly because such a thought would refer to the thought entertained in t1, and so, would not be self-verifying.

Burge replies to this formulation of the incompatibilist challenge by denying (P3'). His fundamental idea is that "memory is fixed by the content of the thinking that it recalls" (Burge, 1998a: 357), an idea developed as follows:

> Memory need not be about a past event or content at all. It can simply link the past thought to the present, by preserving it. Such cases involve a particular type and function of memory –

> preservative memory – which preserves propositional contents and attitudes toward them, rather than *referring* to objects, attitudes, contents, images, or events. (Burge, 1998a: 357)
>
> In the memory case, the content and referent of the remembered material is not distinct from that of the antecedent thought content, which in ordinary that-clause-type self-attributions is both thought and referred to. (Burge, 1998a: 359)

The crucial point to Burge's defense is the differentiation between *preservative memory* and *memory by discrimination*, and the insistence that the first is also essential to understanding such a phenomenon. This difference corresponds to the double interpretation of the question about whether an individual "knows what he was thinking yesterday" (Burge, 1998a: 362). If S relies upon memory to identify a past object or event – including a past thought – S will be subject to error; nevertheless, if S thought yesterday that twaluminum is beside him, he is in a position, relying on preservative memory, to remember what he thought then (Burge, 1998a: 367, footnote). The difference between both situations lies in the difference between a content being fixed in a past thought that is recalled in the present and the other situation where the present thought refers to a past one. The latter idea is not what Burge means by the function of preservative memory.

Following this reasoning, (P3') is clearly false on Burge's account. Just in case one had discrimination in mind, one could infer that in slow-switching cases S does not know what he was thinking yesterday because he is unable to discriminate between two seemingly relevant possibilities (Burge, 1998a: 362). But, in preservative knowledge S knows that p at time t2. As Burge puts it, "[p]reservative memory normally retains the content and attitude commitments of earlier thinkings, through causal connections to the past thinkings" (Burge, 1998a: 357).

Another important point insisted upon by Burge is how the second premise must be defended –(P2'): in the cases presented S

does not forget anything. Boghossian supposes that when switches take place, one set of concepts is displaced by the other one, while Burge does not support such extravagance[13]. To Burge, by no means S forgets one set of concepts when they are replaced by their counterparts. His proposal is to think about a scenario where "the individual has, without realizing it, both the original concept and a new concept after slow-switching" (Burge, 1998a: 368). In this sense, (P2') is completely defended by Burge, because the original beliefs are not forgotten, even if the subject can fail to access them in certain circumstances. Burge stresses that:

> Displacement was never part of the switching cases, at least in my understanding of them. Cohabitation was always the assumed case. I did not and do not consider the displacement model (as a general model for switching cases) a plausible account. (Burge, 1998a: 364-5, footnote 13)

Burge offers another criticism of Boghossian's argument in the sense that if displacement is behind such an argument, it seems that (P2') is mistaken. He says "if one loses a concept when it is replaced by a new one, and for that reason one has no access to beliefs one once had, one may lose knowledge one once had" (Burge, 1998a: 369). In this sense, the argument seems to fail in Boghossian's very framework.[14]

---

13 Actually Boghossian acknowledges both options of reading slow-switching cases but he finally endorses the "displacement" model.

14 An alternative response to incompatibilism could be inspired by Ludlow's comments (1995b). He insists on the falsity of (P1'). Ludlow claims that "Boghossian is correct in asserting that I do not know at t2 what I knew at t1, but he is incorrect in supposing that "the only explanation" for this is that I "never knew" my thoughts in the first place" (Ludlow, 1995b: 310). According to Ludlow, "It is entirely consistent with the social externalist view of memory that I forgot nothing, but that the contents of my memories have nonetheless shifted. Indeed, this is not only possible according to social externalism, but given the prevalence of slow-switching it should be a rather common state of affairs" (Ludlow, 1995b: 310).

Ludlow's position has some serious problems. He claims that Boghossian's

According to this, the argument is unlikely sustained: on the one hand, if we insist that there is something like preservative memory, we should deny (P3'); on the other hand, if we insist on the very Boghossian's view, that Oscar has his mental content replaced according to each world where he is located, and because of that, he doesn't know if he thinks about water or twin water, (P2') is in danger. This kind of dilemma could be used to refuse such an argument as constituting a real risk to compatibilism.

---

argument depends on an individualistic assumption about the nature of memory. According to Ludlow, the contents of our memories are subject to the same external conditions as every mental content is, and he understands by this that those external conditions must be the current ones. One of his serious problems is that he has a misconception of what is the most appropriate externalist account of memory.

Ludlow maintains that social externalism "is bound to say that the content of a memory is fixed at the time recollection takes place" (Ludlow, 1995b: 308). Otherwise, he says, one must accept that those contents are totally inert to all environment changes, and this seems to be contrary to externalism. Ludlow sees a problem in considering memory content somehow as "frozen up" to some later moment of recollection coexisting with the thesis that such contents are fixed by our social environment (Ludlow, 1995b: 309).

However, in the case of mental content of memories, there is no problem at all in accepting that their individuation factors held in the past. After all, memory is about the past. It is about recalling a past thought, with its past content, no matter what the current situation is. There is nothing problematic in being externalist and accepting it. The point is that externalism is not committed to the idea that mental contents are fixed by current external factors, but instead that such contents are individuated by external factors. And the history of this dependence relation matters here.

It seems that Ludlow's solution, in order to solve the incompatibilist challenge, turns the phenomena of memory into a completely empty and absurd faculty. For memory is about to recall the same thoughts one had entertained in some circumstance in the past. Once content of memory is taken to be individuated by current factors, memory no longer can do what it was supposed to do (Ludlow, 1996: 314). In this sense, the immediate conclusion would be that one can seldom remember the thought one had earlier. And this is also quite unacceptable.

Ludlow doesn't seem to have many resources to avoid such criticisms. And it seems clear that his mistake is to suppose that externalism must take memory as he describes. What he conceives as memory cannot, after all, be classified as such.

2. Thought Experiments and Compatibilism: Tyler Burge and Donald Davidson

So far we have seen two incompatibilist instances suggested by the thought experiment of switching cases. However, we could go back and question about the reason why such contexts have received a privileged role within the philosophical debate. It is not obvious how this kind of thought experiment has anything to do with testing the idea that "if externalism obtains, then privileged self-knowledge doesn't", unless one has already supposed the problem to be the following: "to understand how we could know some of our mental events in a direct, non-empirical manner, when those events depend for their identities on our relations to the environment" (Burge, 1988: 650)[15]. It is by translating the incompatibilist risks in those terms that it becomes clear how slow-switching cases match this puzzling intuition.

Burge indicates that even in an extreme scenario, where one's own thoughts are individuated by external factors which are unknown to the subject of the experiment, such a subject is still able to know some of her thoughts in a privileged way. In this sense, Burge highlights compatibilism by reasoning about a scenario where privileged self-knowledge is not undermined by a failure in one's knowledge of one's environment.

Those conditions are, in a sense, very similar to the conditions that hold in the Cartesian demon thought experiment, where one could have direct and non-empirical self-knowledge while doubting completely the existence of a physical world. Skepticism is not in question here, but it is important to notice that part of Burge's strategy lies in insisting that the inference from the Cartesian account on self-knowledge to individualism is misleading (Burge, 1988: 651-52). In fact, Burge makes it clear that part of his aims

15 It was actually Burge who first indicated this sort of puzzle.

is to sustain a "restricted Cartesian conception of self-knowledge" (Burge, 1988: 649).

If this is so, if part of the compatibilist task lies in deconstructing the connection between individualism and this kind of approach to self-knowledge, maybe we could dispense with the use of a thought experiment[16]. It would be enough to consider what John Heil suggests: "If the contents of one's thoughts were determined entirely by the state of one's brain, why should this fact alone make our access to them any less indirect or difficult?" (Heil, 1988: 247). It seems clear that if one had Cartesianism in mind, internalism and total access to one's own mind were to be blended in one and the same position. But if one departs from the question about the nature and the individuation of mental contents, internalism is not equal to total access to one's own mind. It is instead a position which defends that one's mental states are to be individuated by internal factors to the head, such as brain states. As Heil (1988: 247) indicates, there is no clear point in saying that just externalist theories of contents could motivate doubts about the possibility of privileged access.

So, it seems that the inference from privileged self-knowledge to individualism could be easily undermined if we realized that internalism can be much wider than Cartesianism. However, there are other questions involved in the compatibilist enterprise. As Burge himself indicates: "It is one thing to point out gaps in inferences from self-knowledge to individualism. It is another to rid oneself of the feeling that there is a puzzle here" (Burge, 1988: 652). And Burge's slow-switching cases deals with another important question: the dependence between kinds of knowledge.

16 Actually, Burge recognises it, saying that such an inference was already showed to fail by Arnauld's comments on Descartes. Nevertheless, Burge suggests that undermining the Cartesian inference still leaves us with the puzzling sensation that there must be something wrong with externalism. I will insist on the step of rejecting this inference by showing that there are other individualist positions that should deal with the same problems externalism is accused of.

It seems that, for Burge, answering a question such as: "Why is our having non-empirical knowledge of our thoughts not impugned by the fact that such thoughts are individuated through relations to an environment that we know only empirically?" (Burge, 1988: 652-53) involves arguing in favor of the independence of kinds of knowledge: self-knowledge and world's knowledge. Taking slow-switching cases to be a good context of discussion seems to localize Burge's position very close to the skeptic's, because instead of considering such cases as abnormal ones, Burge prefers to state that self-knowledge is left untouched while one can be completely ignorant about one's own environment[17].

Davidson, who agrees with Burge in defending compatibilism, "[does] not consider Burge's thought experiments as persuasive as he does" (Davidson, 1988a: 665), maybe because he defends that self-knowledge and knowledge of the world are interdependent. Furthermore, they are also interdependent of knowledge of other minds. As Burge also does, Davidson accepts the following ideas: "that the contents of our thoughts are individuated in part on the basis of external factors of which the thinker may be ignorant, and that thinkers are authoritative with respect to the contents of their thoughts" (Davidson, 1988a: 664). But it seems that, for Davidson, the concern about how we can know our thoughts without knowing the world in advance must be dissolved instead of answered. The

17 I have, however, exaggerated Burge's position here. The subject's ignorance is localized. Actually, the only statement Burge commits himself to is that there is not an easy answer to the skeptic through externalism. However, once Burge's externalism demands that those proper connections between mind and world must have occurred in order to one's possession of thoughts, he seems to avoid some skeptical worries. In fact, he would not accept general skeptical scenarios so easily; he would first ask the skeptic to explain how the deluded individual has acquired his concepts; and second, if the answer was that the demon has induced him, Burge would argue that the demon would probably have had connections with the world. Nevertheless, I will insist that slow-switching cases seem to share some similarities with Cartesian thought experiments.

point is that we need world knowledge (as well as knowledge of other minds) in order to know our thoughts, but also the other way round. So, there is no question about priority here, nor a problem about world information being required as an enabling condition to knowledge about oneself. One's self-knowledge is also required in order to know the world.

Davidson states that the basic reason for him to hold compatibilism is that "what determines the contents of thoughts also determines what the thinker thinks the contents are" (Davidson, 1988a: 664). In a sense, this totally coincides with Burge's position, yet it seems weaker than appealing to a range of self-verifying thoughts.

As already indicated, the most widely accepted compatibilist answer has been based on the fact that the second-order thought somehow involves the first-order thought, which is individuated externalistically. Such an element has been used to show that in fact "there is no *special* problem for the achievement of self-knowledge in the fact that my first-order thinking is subject to an externalist dependence thesis" (Davies, 2000: 391), although this fact clearly does not explain by itself how it is that my second-order thought amounts to knowledge.

Davidson's compatibilism makes use of such an element. However it is important to recognize that not only his theses about radical interpretation but also the one about the interdependence between the three kinds of knowledge play a decisive role both in Davidson's externalism and in Davidson's compatibilism[18]. As Heil (1988: 247) points out, Davidson's compatibilism indicates that the problem lies not in how externalism deals with privileged self-knowledge, but in a problematic "picture of mind", that needs to be solved[i]. It is a picture where "beliefs about the contents of one's mental states are taken to be based on inward glimpses of those

18 See Davidson (1973) and Davidson (1991).

states or on the grasping of particular entities (contents, perhaps, or propositions, or sentences in mentalese)" (Heil, 1988: 247). Davidson recommends that we abandon the idea that knowledge of mental contents requires our inwardly perceiving in such a way. Once we do so, we remove at least one of the reasons for assuming that externalism undermines privileged access.

This picture of mind is not maintained by Burge either, but there are some important remaining differences between both compatibilisms which seem to refer back to the dependence or independence between kinds of knowledge. Once Burge doesn't see a problem with stating independency, slow-switching cases gain more interest to him than to someone like Davidson, who doesn't see it as a good solution.

Considering the thought experiment as such, Burge seems to provide a consistent compatibilist answer when he maintains his externalist view while appealing to the characteristics of basic self-knowledge. If the question was about the possibility of finding privileged self-knowledge in an externalist framework by offering a range of cases where the answer is positive, Burge reaches a reasonable compatibilist solution.

However, I have suggested that compatibilism could be maintained without giving an answer to such cases. It also seems to be dispensable to insist on Cartesian intuitions in order to talk about privileged self-knowledge. Burge is sympathetic to a restricted Cartesian approach to self-knowledge. But, if by 'restricted' Burge means that just a part of our self-knowledge is acquired in a direct and non-empirical way, there is no need at all to insist on the label 'Cartesian'. A restricted thesis does not seem to be a Cartesian thesis anymore, especially considering that the second-order beliefs partly inherit their content from externalistically individuated beliefs.

This suggests that there are two available paths for the compatibilist to deal with switching cases: to search for an answer to the proposed challenge while maintaining its initial conditions,

as Burge seems to do, or merely solving it, as it seems Davidson does. However, neither of those paths seems to be enough to the establishment of compatibilism, once there is a second context of discussion. A context that would remain intact, even if all the possible problems arisen with the thought experiment are solved: the *reductio ad absurdum* arguments.

3. *Reductio Ad Absurdum* of Compatibilism

The second context of discussion where compatibilism has been tested was initially indicated by McKinsey (1991), but has acquired several formulations, such as Boghossian's:

> Let's suppose that Oscar [...] is a compatibilist. I claim that Oscar is in a position to argue, purely a priori, as follows:
>
> [P1"] If I have the concept water, then water exists.
>
> [P2"] I have the concept water.
>
> Therefore,
>
> [C3"] Water exists. (Boghossian, 1998: 202)

According to Boghossian, (P1") is reached non-empirically by philosophical arguments that sustain externalism while (P2") constitutes Oscar's privileged self-knowledge. Therefore, (C3") could be concluded also by a non-empirical way. And this is the element used against compatibilism: to know a fact of the world, such as the fact that water exists, by a non-empirical manner would be something absurd.

There are several available strategies in order to avoid the alleged incompatibilist result. We could enumerate them as follows: 1. To refuse one of the premises; 2. To defend that the conclusion is not indisputably unacceptable; and 3. to defend that the argument, although being a valid one, has problems that are revealed in terms of epistemic warrants of its elements and how they are related to

each other.

The second strategy is emblematically defended by Sarah Sawyer (1998)[19], who argues that "inferences from introspective knowledge to empirical knowledge are not to be seen as intrinsically unacceptable" (Sawyer, 1998: 528). To consider them as such would constitute a dogma, if our starting point is already externalist. There is nothing epistemically wrong with the argument (Sawyer, 2006). Yet it would be necessary to understand that, for an externalist, to know the world through self-knowledge is not too much to ask, because the concepts of this realm are not themselves unconnected with the world. In order to acquire a concept, a causal connection between the world and my mind is necessary (Sawyer, 1998).

The third strategy has gained a very interesting dimension and, in fact, could be developed under different sub-strategies[20]. Wright's and Davies's analyses represent important strategies within this group. Although they maintain relevant differences between their approaches, both of them indicate that in the argument in question the epistemic warrant of the premises is not transferred to the conclusion. According to Wright (2003), despite the above argument being a valid one, it is not a cogent argument, because the premise's justification seems to require prior epistemic warrant of the conclusion. In this way, the argument would lack the distinguishing feature of leading someone to learn the truth by the justification of the premises, which is the fundamental characteristic of a cogent argument (Wright, 2003: 57).[ii]

Although the above strategies establish important paths in order to deal with the incompatibilist challenge, it seems that the first one would deserve more of our attention because it concerns the very commitments of a compatibilist. On the one hand, it analyzes

19 Warfield (1998) has a similar strategy.

20 See Sawyer (2006) for an overview of the available options. In such a context, the incompatibilist argument is usually taken as a particular instance of a type-argument, of which Moore's proof of the external world (1939) is another case.

what externalism would enable us to know, and on the other hand, what kind of self-knowledge we would have. If our fundamental matter was the incompatibilist discussion, to maintain the argument is a serious mistake, if it is built upon misleading premises.

In the forthcoming lines I shall defend that the argument depends on a misleading conception of externalism.[iii] If, as Sawyer herself points out, "The example is obviously problematic, since no reasonable form of externalism would support the linking conditional stated in [P1"][21]" (Sawyer, 2006: 147), it seems that the urgency lies in deconstructing the argument in that direction.

Boghossian (1998) anticipates two possible ways of rejecting P1": 1. that water would not be required for the acquisition of the concept of water; or 2. that water is required for the acquisition of the concept of water, but this fact could not be known *a priori*. Boghossian argues that such possibilities are easily ruled out, therefore giving rise to incompatibilism. However, Burge (2003b) argues against it, insisting on (1) while Goldberg (2003) insists on (2). Burge claims that:

> Despite its extreme schematic character, this principle [P1"][22] –or any instance of it– is false. As I pointed out in "Other Bodies"; water need not exist in an individual's environment in order for the individual to think that water is such and such. (Burge, 2003b: 262)

Burge suggests that "if one is sufficiently precise, one could introduce a 'natural kind' notion, like water without having had any causal contact with instances of it" (1982: 98, footnote 18). He reminds us that some sciences such as chemistry have indeed anticipated some

21 [P1"] replaces W2 for the sake of text's coherence. In the original text, W2 is the following premise: "If I think that water is wet, then there is water in my environment" (Sawyer, 2006: 147).

22 Burge refers to the following principle:
"WaterDeep Necessarily, for all x, if x is thinking that water is wet then x is (or has been) embedded in such-and-such ways in an environment that contains samples of water" (Burge, 2003b: 262).

natural kinds before their discovery in nature. Externalism doesn't need to deny such a fact.

More than that, Burge adds that an individual or a community could have been mistakenly thinking that there was something such as water. And the point is that if this mistake was discovered, the concept would not be completely emptied (Burge 1982: 97). Burge insists that:

> As I previously indicated, I think that Adam's having attitudes whose contents involve the notion of water does not entail the existence of water. If by some wild communal illusion, no one had ever really seen a relevant liquid in the lakes and rivers, or had drunk such a liquid, there might still be enough in the community's talk to distinguish the notion of water from that of twater and from other candidate notions. We would still have our chemical analyses, despite the illusoriness of their object. [...] I think that Adam's having the relevant attitudes probably does not entail the existence of other speakers. Prima facie, at least, it would seem that if he did interact with water and held a few elementary true beliefs about it, we would have enough to explain how he acquired the notion of water. What seems incredible is to suppose that Adam, in his relative ignorance and indifference about the nature of water, holds beliefs whose contents involve the notion, even though neither water nor communal cohorts exist. (Burge, 1982: 98)

Goldberg (2003), on the other hand, insists that the problem with (P1") lies elsewhere. Regarding the argument as formulated above, water is indeed a necessary condition for the possession of the concept of water, but such a fact could not be known a priori. He claims that:

> The upshot is that McKinsey-style arguments, which would have us conclude [...] that I can know a priori that e.g. water exists, fail, for assuming that all statements expressing metaphysical dependencies between their designata are knowable a priori (…)
>
> Precisely not, since the metaphysical dependence of [<u>water</u>] on the existence of water (H_2O) itself depends on the identification of

water with H_2O. (Goldberg, 2003: 40-1)

Although the latter may be questionable (that the metaphysical dependence of water on the existence of water (H_2O) itself depends on the identification of water with H_2O) it seems fair to accept that the question about whether water is necessary for the acquisition of the concept of water is a matter of empirical knowledge. In that sense, this indicates an alternative route to reject (P1")[iv]. Nevertheless, such a route seemingly has a very narrow application. It could only be applied to the kind of externalism deduced from Putnam's works (1979) and which is the base of the incompatibilist argument proposed by Boghossian (1998).

In Putnam's context (1975), if the external trait of Oscar's mental states is explained by the fact that water has caused such a thought, it seems that (P1") would be available to an externalist[23]. If Oscar has the concept of water, and he is an externalist, the fact that water exists would be available for him. If it is discovered that he was wrong, that in fact water doesn't exist, what Oscar had was not a concept but a pseudo-concept instead. It is here where Goldberg's criticism has an application. An externalist would be able to reach (P1") because, besides knowing the philosophical arguments that have led him to (P1"), he had knowledge of the world, in this case, about the constitution of water. (P1") could be the Putnamian lesson in 1975, and so it could be attacked following Goldberg's criticism, but it should not be blended with the several available externalist positions.

This represents a third route to indicate that (P1") has to deal with serious objections. (P1") is based on an externalist position which is neither the unique nor the prevailing one. An externalist

23 That Putnam is committed to (P1") seems to be connected with his conception of meaning as being composed by stereotypes plus reference. However, in what follows, I shall develop another line of argumentation. I will indicate one possible interpretation of Putnam's position that sees him as committed to the idea that reference points to the sufficient and necessary cause of one's thoughts.

does not need to sustain that the external trait of my concepts is due to the supposed fact that the related objects of my concepts have caused them. At least, not in the atomist way as it seems to be assumed by the incompatibilist argument.

An externalist position emphasizes that the mind is constituted by the external to our skin because we interact with our world and with our community. Some positions explain such interaction appealing to causality, explaining, for instance, that our mind is constituted by what is external to our skin through causal relations between oneself, one's fellows and one's world. Others prefer to explain such interaction by appealing to our linguistic abilities and to the notion of objectivity, explaining that our mind is constituted by what is external to our skin because the base of our mental realm is constituted by a reasonable range of knowledge. However, it seems that just a few positions would sustain that it is possible to deduce, from each of our concepts, a correspondent object to which we could refer in order to explain the history of the acquisition of that concept. In an externalist framework, a mental holism seems to have much more space than an atomism. An atomism seems to require, in fact, that some of our mental contents need to be identified in an internalist manner.

Let's consider an atomist position in which each of our concepts should correspond to an item of the world. Such a position is clearly problematic, once we have concepts without correspondents in the world, such as the well-known example of the unicorn. How should those concepts be individuated? If we follow the atomist line of reasoning, they might be individuated by an internalist manner. If there is no such a correspondent in the world, we might explain them as being pseudo-concepts or in terms of something internal to our heads. In this way, when an atomist considers himself able to explain the external character of some of our concepts, he only achieves this by maintaining another broad group of concepts individuated by an internalist way.

However, the aim of several externalist positions is to sustain that at least part of all our mental contents are constituted in an externalist manner. The very idea of internalistically individuated contents – the narrow contents – has been under attack under the accusation of being untenable notions. This has also been an important criticism directed to Putnam's position in 1975, when his externalism was sustained upon the price of the necessity of narrow contents.

In his way, although Goldberg's criticism seems to be a good one, it has its own scope diminished because he contemplates a condition only sustained by a specific and controversial kind of externalism. Because of that, it seems that the most reasonable thing to do would be to insist that (P1") is not an externalist consequence, by the reasons indicated above as well by Burge's reasoning. In doing so, the argument in question here could not serve as the basis of an incompatibilist attack.

Conclusion

Despite the variety of questions and arguments treated in the text, I have defended a general thesis in the following terms: the incompatibilist challenges introduced in the literature under the form of slow-switching cases and *reductio ad absurdum* arguments represent a real objection to compatibilism only under very specific conditions: if we assume a specific account of self-knowledge in the first case and if we assume a specific approach to externalism in the second case. When we move on to other approaches, the incompatibilist risk is solved.

In the first part, I have treated the slow-switching cases such as exposed by Burge (1988). I've discussed two incompatibilist arguments based on Boghossian's comments (1989a) and some of the ways they could be answered.

The first interpretation of Boghossian's comments has pointed

out that the switched subject could not have knowledge of her own thoughts since she was unable to discriminate between water and twin water thoughts. The second argument has driven a criticism over compatibilism appealing to questions about memory. Taking Burge's framework about self-knowledge as the starting point, the argument has indicated that even though the subject of the experiment could know her own thoughts in a privileged manner at the moment she was thinking them, she would be unable to remember them later on.

Considering Burge's comments on that question, the latter argument would lead us to a kind of impasse: on the one hand, if we accept that there is something like preservative memory, we would have to disregard (P3'). On the other hand, if we maintain (P3'), we would have to abandon (P2'). Therefore, I have defended that such an argument could not provide a basis for an incompatibilist position.

Regarding the former argument, I have defended that it could only be used as a support for incompatibilism if we insisted on the following view about self-knowledge: all that deserves the label of self-knowledge might be potentially available to be known in a direct and non-empirical manner. However, I've argued that such a vision about self-knowledge would be so problematic as to sustain that there is no parcel of privileged self-knowledge. A fair account of self-knowledge should give rise to the privileged kind of acquisition as well as to the indirect and empirical method by means of which we know part of our minds.

In the intermediate part, I have raised the question about the role the thought experiment occupies in the attempt to defend compatibilism. I've discussed two compatibilist frameworks, Burge's and Davidson's. While Burge proposes the experiment, offering a compatibilist answer to it, Davidson would instead tend to solve it. I have suggested that Davidson is also able to provide a compatibilist framework following a different route, one that dispenses with Burge's commitments, such as the reference to basic self-knowledge

and his supposed commitment to the thesis of independence between kinds of knowledge. I have also referred to the idea that the external character of the second-order thoughts is due to the fact that they somehow embed the first-order thoughts, which are externalistically individuated. In this way, there would not be a special problem about how to explain the privileged acquisition of self-knowledge while externalism is in place. The vertigo of puzzle, however, should be cured together with the dissolution of the Cartesian approach to self-knowledge.

The last part of the text has treated the second group of incompatibilist challenges, the *reductio ad absurdum* argument, as exposed by Boghossian (1998). I have defended that if such an argument was based on a misleading conception of externalism, it should be rejected as a good support for incompatibilism. Taking into account Burge's (2003b) and Goldberg's (2003) reasons to reject (P1"), it was defended that Putnam's externalism (1975) would be the only instance where (P1") might have space. That is, disregarding Goldberg's criticism (2003), the only kind of externalism that could have (P1") as a consequence of its theses would be an externalism of Putnam's type. It was argued that externalism, in general, neither needs nor is committed to the implication involved in (P1"). To sustain that our thoughts are identified in relation to external factors doesn't give us the right to infer the existence of a supposed correspondent of the mental content in the world. If there is in the world a referent of a particular concept, and if it had some importance in the acquisition of the thought in question, this acquisition would not be independent of the community or even of the very individual.

Bibliographical References

Boghossian, P. 1989a, "Content and Self-Knowledge", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI

Publications, pp. 149-73.
Boghossian, P. 1998, "What the Externalist Can Know 'A Priori'", *Philosophical Issues* 9, pp. 197-211.
Brueckner, A. 1997, "Externalism and Memory", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 319-31.
Burge, T. 1982, "Other Bodies", in T. Burge 2007, *Foundations of Mind*, Oxford, Oxford University Press, pp. 82-99.
Burge, T. 1986, "Cartesian Error and the Objectivity of Perception", in T. Burge 2007, *Foundations of Mind*, Oxford, Oxford University Press, pp. 192-207.
Burge, T. 1988, "Individualism and Self-Knowledge", *The Journal of Philosophy* 85 (11), pp. 649-63.
Burge, T. 1998a, "Memory and Self-knowledge", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 351-70.
Burge, T. 2003b, "Replies from Tyler Burge", in M. J. Frápolli & Romero, E. (eds.) 2003, *Meaning, Basic Self-knowledge, and Mind*, Stanford, CSLI Publications, pp. 243-96.
Burge, T. 2003/2006, "Descartes on Anti-individualism", in T. Burge 2007, *Foundations of Mind*, Oxford, Oxford University Press, pp. 420-39.
Davidson, D. 1973, "Radical Interpretation", in D. Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 125-39.
Davidson, D. 1987, "Knowing One's Own Mind", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 15-38.
Davidson, D. 1988a, "Reply to Burge", *The Journal of Philosophy* 85 (11), pp. 664-65.
Davidson, D. 1991, "Three Varieties of Knowledge", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 205-20.
Davies, M. 2000, "Externalism and Armchair Knowledge", in P. Boghossian & C. Peacocke (eds.) 2000, *New Essays on the A Priori*, Oxford, Oxford University Press, pp. 384-414.
Davies, M. 2003, "Externalism, Self-Knowledge and Transmission of Warrant", in M. J. Frápolli & E. Romero (eds.) 2003, *Meaning, Basic Self-knowledge, and Mind*: *Essays on Tyler Burge*, Stanford, CSLI

Publications, pp. 105-30.
Goldberg, S. 2003, "On Our Alleged A Priori Knowledge that Water Exists", *Analysis* 63 (1), pp. 38-41.
Heil, J. 1988, "Privileged Access", *Mind* 97 (386), pp. 238-51.
Ludlow, P. 1995a, "Externalism, Self-Knowledge, and the Prevalence of Slow Switching", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 225-30.
Ludlow, P. 1995b, "Social Externalism, Self-Knowledge, and Memory", P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 307-10.
Ludlow, P. & Martin, N. (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications.
McKinsey, M. 1991, "Anti-Individualism and Privileged Access", *Analysis* 51, pp. 9-16.
Putnam, H. 1975, "The Meaning of 'Meaning", in H. Putnam 1975, *Mind, Language and Reality,* Philosophical Papers, vol. 2, Cambridge, Cambridge University Press, pp. 215-71.
Sawyer, S. 1998, "Privileged Access to the World", *Australasian Journal of Philosophy* 76 (4), pp. 523-33.
Sawyer, S. 2002, "In Defence of Burge's Thesis", *Philosophical Studies* 107, pp. 109-28.
Sawyer, S. 2006, "Externalism, Apriority and Transmission of Warrant", in T. Marvan (ed.) 2006, *What Determines Content? The Internalism / Externalism Dispute*, Cambridge, Cambridge Scholars Press, pp. 142-53.
Warfield, T. 1992, "Privileged Self-Knowledge and Externalism are Compatible", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 215-21.
Warfield, T. 1997, "Externalism, Privileged Self-knowledge, and the Irrelevance of Slow Switching", in P. Ludlow & N. Martin (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications, pp. 231-37.
Warfield, T. 1998, "A Priori Knowledge of the World: knowing the world by knowing our minds", *Philosophical Studies* 92, pp. 127-47.
Wright, C. 2000, "Cogency and Question-Begging: Some Reflections on McKinsey's Paradox and Putnam's Proof", *Philosophical Issues* 10, pp. 140-63.
Wright, C. 2003, "Some Reflections on the Acquisition of Warrant by

Inference", in S. Nuccetelli (ed.) 2003, *New Essays on Semantic Externalism and Self-Knowledge*, Cambridge, Mass., MIT Press pp. 57-77.

i Such a line of argumentation is found in Davidson (1987).

ii Other examples that somehow follow this strategy can be found in Beebee (2001) and in Lewis (1996). Beebee (2001) corroborates the main traits of Davies' account, but she emphasizes the phenomenon of 'begging the question' instead of 'failure of transfer of warrant'. In the case of Moore's (1939) proof of the external world, for example, there is no warrant at all that fails to be transferred from the first premise (I have two hands) to the conclusion, if the audience is sceptical. Moore's proof "fails to convince in the sense that it cannot persuade rational doubters of the conclusion, not because doubters may not transfer warrant from premises to conclusion, but because rational doubters will not accept that they have warrant for the premises in the first place" (Beebee, 2001: 359). According to Lewis (1996), this sort of arguments suffers a context shift from one premise to the other; still considering Moore's proof, while we are fully justified in stating "I have two hands" in standard situations, if the situation envisaged is one where questioning the existence of the external world makes sense, then such a possibility can no longer be ignored and we cannot say anymore that we know we have two hands.

iii The first strategy of rejecting one of the premises of Boghossian's arguments is also found in Corbí (1998), who attacks the second premise. He argues that "we cannot know 'a priori' that I master a certain concept, since it seems clear that it could occur that I would wrongly believe that I master a certain concept" (Corbí, 1998: 232). "Moreover, the idea of 'a priori' knowledge seems to comprise an internalist element: knowing that p 'a priori' involves my being able to provide certain reason to justify my believing that p –reason that, indeed, should not mention any particular facts of the world […] But, it seems that this cannot be done […] It sounds then that my knowledge of the concepts I master is not 'a priori', even if it is special" (Corbí, 1998: 235).

iv See Moya (1998) for a detailed discussion of the defense that (P1") cannot be known *a priori*.

5 Interpretando la Paradoja de Moore: la irracionalidad de una oración mooreana[1]

Cristina Borgoni. Published in *Theoria*, vol. 23/2, num. 62, pp. 145-161, 2008.

ABSTRACT
This paper offers an interpretation of Moore's Paradox that emphasizes its relevance for our understanding of rationality and linguistic interpretation. The sentences that originate the paradox do not need to be thought of in terms of the absence of a contradiction, but in terms of absence of rationality, where rationality is understood as a broader notion than coherence and logical consistency. This is defended through three theses, two of which stem from the dominant (but insufficient) approaches to the paradox: Moore's, Wittgenstein's and Shoemaker's.

KEYWORDS: Moore's paradox, rationality, radical interpretation, personal unity, externalism and first person perspective.

RESUMEN
Este trabajo ofrece una lectura de la Paradoja de Moore que pone énfasis

1 Versiones anteriores de este trabajo fueron presentadas y discutidas en la Universidad de Granada, en marzo de 2007, en el marco del V Seminario de Trabajos en Construcción, y en Barcelona, en septiembre de 2007, en el marco del V congreso de la SEFA. Agradezco el interés y los comentarios allí recibidos. Agradezco especialmente a Manuel de Pinedo García, a Neftalí Villanueva Fernández, a Mar Muriana López, a Miguel Ángel Pérez y a los dos evaluadores anónimos.

en su relevancia para nuestra comprensión de la racionalidad y de la interpretación lingüística. Mantiene que las oraciones que dan origen a la paradoja no necesitan entenderse en términos de ausencia de una contradicción, sino más bien en términos de ausencia de racionalidad, entendida esta como un término más amplio que el de coherencia y consistencia lógica. Se defenderá tal posición por medio de tres tesis, dos de las cuales se derivan de los enfoques dominantes (aunque insuficientes) a la paradoja: el de Moore, el de Wittgenstein y el de Shoemaker.

PALAVRAS CLAVE: paradoja de Moore, racionalidad, interpretación radical, unidad personal, externismo y la perspectiva de primera persona.

Introducción

Imaginemos que llueve y que estamos una amiga y yo frente a una ventana donde se ve claramente tal situación. Ella dice: "Está lloviendo". Yo la escucho y entiendo que quiere decir algo respecto al tiempo; quiere, por ejemplo, llamarme la atención sobre un evento que estamos presenciando y sobre el que, de alguna manera, podemos pensar. En seguida afirma, "pero no creo que esté lloviendo". ¿Qué puedo decir en respuesta? Por un lado, soy capaz de ver que llueve y sé que se equivoca cuando afirma que no cree en ello. Pero, por otra parte, al haber dicho "llueve" al principio, me deja sin espacio para darle cualquier razón que la convenza de lo contrario con relación a la segunda parte de su enunciado. Parece que solamente me resta insistirle "¿Qué?" A pesar de tal impase, tengo que aceptar que mi amiga ha proferido una oración gramaticalmente correcta.

La discusión sobre la Paradoja de Moore se establece en contextos similares a este, donde existe un enunciado que, en principio no tiene ningún problema formal, pero resulta ser un enunciado absurdo. Tradicionalmente, la paradoja se caracteriza por un consenso sobre el carácter absurdo de tales oraciones, a la vez que

no existe ninguna contradicción en la oración en sí misma. Por esto, la discusión se convierte en una búsqueda de lo que podría ser una contradicción disfrazada.

En este texto buscaré ofrecer una lectura de la paradoja que ponga énfasis en su relevancia para nuestra comprensión de la racionalidad y de la interpretación lingüística. Defenderé que tales oraciones paradójicas no necesitan entenderse en términos de ausencia de una contradicción sino más bien, en términos de ausencia de racionalidad. Lo haré por medio de la defensa de tres tesis, dos de las cuales se derivan de los enfoques dominantes (para mí, insuficientes) a la paradoja: el de Moore, el de Wittgenstein y el de Shoemaker[2]. Con esto, intentaré mostrar que tal paradoja no necesita verse como un problema encerrado en sí mismo, sino que el camino para solucionarla pasa por cuestiones importantes concernientes a cómo funciona nuestra interpretación lingüística, a la auto-atribución de conocimiento y al carácter externo de lo mental.

Mi estrategia será: primero, estimular el carácter paradójico de las oraciones en juego, luego presentar las tres explicaciones dominantes acerca de la paradoja y por último, exponer mi propuesta y los temas que están involucrados en ella.

1. Qué hay de paradójico en las "oraciones mooreanas"

Consideremos las siguientes oraciones:

i. *p*
ii. Creo que *p*

2 Agradezco a un evaluador anónimo el hacerme ver la importancia de la solución de Hintikka (1962). Sin embargo, he optado por hacer referencias a ella en distintas partes de este trabajo en lugar de tratarla como una cuarta lectura de la paradoja. Como señalaré, las ideas de Hintikka tienen aspectos en común con algunas de las soluciones que serán estudiadas, incluyendo la que yo propondré.

iii. No creo que *p*

iv. Creo que no *p*

como expresión de situaciones en el mundo. Puede pasar que lo que (i) expresa ocurra conjuntamente con cada uno de los casos expresados por las tres siguientes oraciones: *p* podría ser el caso y además de esto, yo podría creer que *p*; otra situación sería aquella en la que *p* fuera el caso y yo ignorara tal cosa; y finalmente, *p* podría ser el caso y a la vez yo podría equivocadamente creer en la negación de *p*. A pesar de la plausibilidad de tales situaciones, enunciar (i) más (iii) o (i) más (iv) como:

1. p, pero no creo que *p* (i & iii)

y

2. p, pero creo que no *p* (i & iv)

produciría, en ambos casos, enunciados muy extraños.

Es aquí donde la "paradoja de Moore" aparece y por esto, a partir de ahora, las oraciones de los tipos (1) y (2) pasarán a llamarse "oraciones mooreanas" (OM). El mundo puede fácilmente aceptar que algo sea el caso sin que yo crea en ello, pero tenemos el deber de recusar las palabras de alguien que enuncia un hecho y a la vez enuncia su creencia en lo contrario. Del mismo modo que debemos rechazar que alguien afirme algo a la vez que afirma la ausencia de su creencia en tal cosa. Dicho de una manera más tradicional, no hay ninguna contradicción entre (i) y (iii) ni tampoco entre (i) y (iv) pero, aún así, cuando tales conjunciones están afirmadas en primera persona se vuelven inaceptables, porque son absurdas.

Según Williams (2006: 227) Moore habría tenido el cuidado de distinguir un absurdo de una paradoja. En el caso de las oraciones mooreanas, lo absurdo sería que yo afirmara tales oraciones. Lo paradójico sería que el absurdo coexistiera con la ausencia de cualquier contradicción en mis palabras.

A lo largo del texto voy a mantener que tales oraciones son, de hecho, inaceptables, porque dejan de responder a un comportamiento racional y no por ser un equívoco lógico. Pero, por

ahora es suficiente con que sigamos delimitando el tipo de oraciones con las que estamos tratando. Es interesante señalar la peculiaridad de (1) y (2) frente a otras oraciones que tienen alguna similitud con una OM, pero no lo son, como las siguientes:

3. *p,* pero ella no cree que *p*

4. *p,* pero yo creía que no *p*

5. *p,* pero sé que no *p*

No hay nada de contradictorio ni tampoco de paradójico en enunciar (3) o (4). Ambas podrían ser incluso verdaderas.

El aire paradójico de "*p,* pero no creo que *p*" en contraste con "*p,* pero ella no cree que *p*" (o incluso "*p,* pero yo creía que no *p*") nos invita a reconocer la diferencia entre puntos de vista de primera y de tercera persona. Aunque (1) y (3) tengan la misma forma lógica (p & $\sim B_S p$), la diferencia entre que una sea paradójica y la otra no, se halla en quién es S; quién es el sujeto que cree en *p*[3]. En (1) hablo de mí misma mientras que en (3) hablo de otra persona. Y, con frecuencia, nos apercibimos de la ocurrencia de determinado evento al tiempo que reconocemos que hay personas que no creen que suceda.

Con respecto a (4), la oración es igualmente familiar. Son muchas las veces en que nos damos cuenta de que algo es el caso aunque creyéramos erróneamente en lo contrario. En este caso, la oración la enuncio yo sobre mí misma, pero también desde una perspectiva de tercera persona. Tengo que verme a mí misma en el pasado y comprenderme como poseedora de una creencia que ya no sostengo. Y para esto parece necesario un cierto alejamiento que se consigue asumiendo una perspectiva de tercera persona sobre mí misma. Si digo (4), al localizar mi creencia en un tiempo distinto,

3 Hintikka (1962: 104) llama la atención sobre tal diferencia y reconoce que el pronombre de primera persona 'yo' juega aquí un papel especial. Sin embargo, sugiere que no solamente sería preferible explicar la paradoja sin añadir criterios que hicieran referencia al mismo, sino que una explicación de este tipo, "en términos de las características propias de los pronombres de primera persona" estaría en principio mal encaminada.

me libro de estar en una situación absurda[4].

La quinta oración es bastante distinta de las anteriores, porque en ella sí podemos encontrar una contradicción clara entre *p* & no *p*. Y esto se debe a la facticidad del conocimiento. En el caso (5) podríamos decir que yo, o quien lo afirmara, estaría cometiendo un equívoco porque estaría afirmando una contradicción, pero no tendríamos problemas en explicar el porqué. Así que, no estaríamos ante una oración paradójica. Si alguien conoce *p*, implica que *p* es el caso. Aquí no hay ningún misterio.[5]

Sería posible establecer otras relaciones entre las oraciones mooreanas y otras que se parecen a ellas pero no lo son. Sin embargo, creo que estas son suficientes para marcar algunas especificidades de una OM, y más que esto, motivar el verlas como oraciones problemáticas. El hecho de que las oraciones mooreanas sean una conjunción entre un enunciado del tipo "*p*" y otro del tipo "S cree que *p*" (o bien S no tiene la creencia en *p*, o bien S cree en lo contrario de *p*), y el hecho de que se afirmen por la primera persona y en tiempo presente, son suficientes para que surja la siguiente cuestión: ¿qué tiene de absurdo una OM? En la próxima sección veremos una variedad de respuestas a tal pregunta.

---

4 Alguien podría contra-argumentar que asumir la perspectiva de tercera persona sobre mí misma no me libra de caer en algún momento en la paradoja de Moore. Tal perspectiva podría tomarse también para hablar sobre mis estados de creencia actuales con la ayuda, por ejemplo, de un psicólogo, y podría así afirmar algo como la oración (1). R. Moran (2001) llama la atención justamente sobre casos como estos en los que podríamos tener una perspectiva de primera persona conflictiva con una de tercera persona sobre nosotros mismos. Estoy totalmente de acuerdo con que casos así podrían darse. Sin embargo, la peculiaridad de (4) está en que el modo como accedo a mí parece estar presente en la propia oración. No hay ningún problema en que yo actualmente crea en p, aunque no haya creído en el pasado.

5 T. Williamson (2000) defiende que oraciones como la (5) pueden entenderse como casos de la paradoja de Moore.

2. Disolviendo la paradoja

Dar una solución a la presente paradoja es descubrir dónde está lo absurdo de una OM. Para esto, se ha sugerido que debemos localizar el fenómeno que se parece a una contradicción en otra parte que no sea la oración en su forma primitiva. Hay, en realidad, muchas propuestas de solución en la literatura, pero, por lo menos tres me parecen especialmente importantes y serán estas las que consideraré: la primera está inspirada por el mismo Moore, la segunda desarrollada por Wittgenstein, y una tercera por Shoemaker. Las analizaré con la intención de buscar las principales intuiciones que subyacen a cada solución.

G. E. Moore

La paradoja de Moore, incluso la invención del propio término 'Moore's Paradox', debe su significado filosófico, tal y como lo conocemos y lo discutimos hoy, a Wittgenstein (1953). Según observa Baldwin (1996: 226), G. E. Moore habría presentado su paradoja en dos trabajos tardíos y en ninguno de ellos la paradoja era el tema principal. La utilización de los enunciados que dieron lugar a ella, servía solamente para destacar la distinción entre lo que una persona afirma y lo que implica al hacerlo.

Es posible, sin embargo, encontrar una solución a este problema inspirada en los textos del propio Moore. En su famoso ejemplo: "Fui a la exposición el jueves pasado, pero no creo que fui"[6] (Moore, 1942: 543), Moore sostiene que mi declaración de que fui a la exposición implica que creo que fui. Y es por esto que sostiene que proferir tal oración sería absurdo.

Una posible interpretación de la idea de que "mi declaración de que fui a la exposición implica que creo que fui" sería mantener

6 Todas las traducciones del inglés son mías a no ser que se indique lo contrario.

que (i) implica lógicamente (ii)[7]. Bajo tal interpretación, si alguien asevera (i), podríamos inferir (ii), permitiéndonos ver la tensión buscada. En el caso de la proferencia de (1) *–p,* pero no creo que *p*– si seguimos tal sentido lógico de implicar, podríamos inferir: (1*) "creo que *p,* pero no creo que *p*", mientras en el caso de (2) *–p,* pero creo que no *p*– tendríamos: (2*) "creo que *p,* pero creo que no *p*".

La diferencia entre las oraciones (1) y (2), del mismo modo que en las resultantes (1*) y (2*), puede, en principio, considerarse como una diferencia demasiado sutil una vez que utilizamos normalmente la oración (iii) "no creo que *p*" con el sentido de (iv) "creo que no *p*". Sin embargo, Williams (1979) sugiere que tal diferencia debe tenerse en consideración porque, en el caso de la paradoja de Moore, se establecen dos problemas donde antes solo se veía uno, de manera tal que su solución debe entenderse desde este doble aspecto.

Según Williams (1979: 142) el absurdo de (1) se distingue del absurdo de (2) de la siguiente manera: en el primer caso sería absurdo que alguien dijera (1) porque lo que se expresa y se afirma conjuntamente, o sea, (1*): una creencia de que *p* y la ausencia de la creencia de que *p*, es *imposible lógicamente*. Mientras que en el caso de (2) lo que se expresa y se afirma conjuntamente, o sea, (2*): una creencia que *p* y una creencia de que no es el caso que *p*, es *inconsistente*.

Seguiré considerando (1) y (2) como casos distintos, pero no me comprometeré con las denominaciones señaladas por Williams para cada caso porque estas solamente tienen sentido si aceptamos esta primera interpretación. Una interpretación que, sin embargo, no parece ser la interpretación más fiel a la posición de propio Moore.

7 Se mantendrá más adelante que tal interpretación no es la interpretación más fiel a la propia posición de Moore. Sin embargo, como se va a discutir también más adelante, la interpretación que sí sería la más fiel no logra una solución satisfactoria en el sentido de encontrar una contradicción en las oraciones en juego. Se considerará inicialmente esta primera interpretación porque además de constituir una lectura posible, parece estar presupuesta por otros filósofos como es el caso de Williams (1979).

Aunque Moore no desarrolle una solución sistemática a la paradoja, se puede entender desde sus escritos que el sentido de la implicación a la cual se refiere no es el de implicación lógica, sino más bien una noción informal y cotidiana[8]. En uno de sus trabajos, Moore (1944: 204, citado por Black, 1952: 26) sostiene que si decimos, por ejemplo, "él no salió" nosotros implicamos que no creemos que él haya salido, aunque esto no haya sido ni afirmado ni pueda seguirse de nada de lo que afirmamos[9]. Según expone, implicamos esto porque llevamos a nuestros oyentes a asumir tal cosa, porque "las personas, en general, no hacen una aserción positiva a no ser que no crean que su opuesto sea verdadero". Por esta razón, la primera interpretación que prometía una solución rápida a la paradoja no se sostiene. En parte porque no es la posición de Moore, pero principalmente porque no parece ser sencillo mantener que haya una implicación lógica entre (i) y (ii). Si uno puede suponer (ii) cuando alguien asevera (i), esto no es lo mismo que sostener que en todo

8 Hay razones de peso para entender que cuando Moore habla de implicación no se refiere a la implicación lógica, sino más bien a la implicación epistémica o pragmática [esto es algo en lo que insiste, por ejemplo, Hintikka (1962) en su lectura de la paradoja]. Max Black (1952), aunque no utilice términos tan precisos como los de la pragmática contemporánea, ofrece una lectura sobre el uso que hace Moore de la noción de inferencia que puede resultar esclarecedor: "De estas observaciones sobre Moore, podemos derivar la siguiente explicación del modo según el cual él [Moore] está usando aquí la palabra 'implica': Suponga (i) el hablante está usando una expresión, *E*, (ii) las personas no usan en general *E* sin que alguna proposición relacionada *pE* sea verdadera, (iii) las personas, al oír el uso de la expresión *E* por el hablante, en general van a creer que *pE* es verdadera, y por fin (iv) el hablante sabe todo esto – entonces si las cuatro condiciones se dan, las palabras del hablante pueden ser tomadas como implicando *pE*" (Black, 1952: 26).

9 Aquí se podría recurrir a nociones pragmáticas como es el caso de la noción de *afirmación* entendida como un acto de habla sujeto a determinadas normas, como por ejemplo, la suposición de que quien afirma algo cree lo que afirma. Pero, de nuevo, es importante señalar que leer a Moore en términos de las categorías de la pragmática contemporánea es una entre varias interpretaciones posibles, ya que él no disponía de tales recursos en sus trabajos.

caso, (i) implica (ii). Y Moore no es ingenuo con respecto a esto.

Nos quedamos entonces con la otra interpretación de la posición de Moore, la que toma el sentido de la implicación entre mi afirmación sobre un evento y mi creencia en tal evento como más débil que el sentido de implicación lógica. Linville & Ring (1991: 295-96), teniendo esto en cuenta, señalan sin embargo que solamente en este segundo sentido (el sentido lógico) tendríamos derecho a indicar la existencia de una contradicción en una OM y así solucionar la paradoja. Una vez que la paradoja es puesta en términos de hallar una contradicción, esta segunda interpretación de Moore se muestra insatisfactoria.[10]

10 Otra posible salida sería tomar una dirección como la que toma Hintikka (1962), que apela a principios de creer y de saber para explicar el carácter paradójico de una OM. De una cierta manera, Hintikka reconoce la importancia del punto señalado por Moore, de que hay un cierto presupuesto en la comunicación de que las personas creen en lo que dicen. Sin embargo, va más allá de tal explicación. Según Hintikka "Moore basa su explicación de [una OM] en el hecho de que en la gran mayoría de los casos nosotros creemos aquello que decimos" (Hintikka, 1962: 129). Pero señala más adelante: "No tengo por qué dar por supuesto que cuando se lleva a cabo un enunciado (*q* por ejemplo) bajo condiciones normales y en un tono de voz natural, haya que presumir que uno cree que *q* es verdadero. Para mis propósitos, basta con que siempre que uno diga que *q*, se suponga que uno puede creer aquello que uno dice (en el sentido de que *q* no ha de ser indefendible)" (Hintikka, 1962: 131). Dice más: "Si estoy en lo cierto, quienquiera que diga [una OM] dice también algo que a la larga le resulta imposible de creer […]. Esta imposibilidad no es sino una consecuencia de las propiedades lógicas de [una OM]" (Hintikka, 1962: 129). Aunque Hintikka parezca guardar la intuición mooreana acerca del problema, va más allá cuando indica que la peculiaridad de una OM se entiende por medio del concepto de *indefendibilidad doxástica* (Hintikka, 1962: 126). En este sentido, tal solución tendría importantes rasgos en común con el enfoque de Shoemaker, que se va a ver adelante, donde se defiende que el problema real estaría en creer en una OM. Pero, de nuevo, no se resumiría a tal posición porque Hintikka es también claro con respecto a tal posibilidad: "Yo no creo que tengan sentido únicamente aquellas formas verbales que puedan ser dichas y creídas (o conocidas) por el hablante. Lo que ocurre es que las formas verbales que no satisfacen esta condición resultan inútiles para la mayoría de los propósitos que uno espera que nuestro lenguaje cumpla" (Hintikka, 1962: 131). Se agradece el comentario de un evaluador

Sin embargo, es posible encontrar virtudes en la contribución de Moore, aunque no solucione completamente el problema. Una cuestión importante que se deduce de su sugerencia, es la idea de que si alguien enuncia *p,* tal persona estaría, la mayoría de las veces, lista para corroborar tal aserción, en el sentido de comprometerse con *p*. Creo que tal enfoque permite hacer justicia a la intuición de que al aseverar *p* lo que se dice no sólo trata sobre el mundo, sino que también trata sobre uno mismo, al menos en el sentido de hacer explícito el propio punto de vista.

Cuando se presentó la paradoja, el hecho de haber una conjunción entre un enunciado supuestamente acerca del mundo y otro supuestamente acerca de la creencia de un sujeto, era lo que nos impedía encontrar una contradicción. Aunque Moore tampoco lo consiga, creo que su esfuerzo es el de aproximar el tópico de los dos enunciados. En el caso de Moore, él señala que las afirmaciones sobre el mundo no están separadas de la existencia de alguien que las afirme. Y si alguien las afirma, entendemos que se está comprometiendo con lo que dice.

L. Wittgenstein

La segunda propuesta de disolución de la paradoja se inspira en Wittgenstein[11]. Tal solución podría entenderse como la estrategia inversa a la primera, aunque por razones bastante distintas. Wittgenstein sugiere que decir "creo que *p*" es solamente otra manera de decir "*p*". En sus palabras: "[E]l enunciado 'Creo que va a llover' tiene un sentido análogo a 'Va a llover', pero 'Entonces creí que iba a llover' no tiene un uso análogo a 'Entonces llovió'" (PI, parte II, sección X).

---

anónimo que indica que la propuesta que desarrollaré más adelante, que explica la peculiaridad de una OM en términos de principios de la racionalidad, podría compartir aspectos con tal propuesta de Hintikka, que toma "los principios de saber y creer" como explicación de su naturaleza absurda.

11 Las propuestas de Heal (1994), Lee (2001) and Linville & Ring (1991), son ejemplos de variantes de la solución wittgensteiniana.

Heal (1994: 20) entiende tal propuesta como la idea de que una persona que aprende un lenguaje está entrenada para usar, en ocasiones, "creo que *p*" como sustituto de "*p*". De ese modo, la supuesta referencia a la creencia de la persona en *p* no añadiría nada, o no diferiría en nada de lo que ya se afirma con *p*. En este sentido, más que invertir la dirección, esta segunda solución se distingue de la primera por explicar dónde está la contradicción sin hacer referencia a lo que una persona implica con sus palabras. El verbo creer, usado en primera persona, podría entenderse aquí prácticamente como un término redundante[12].

Aplicando tal razonamiento a la paradoja, tenemos que una OM sería sencillamente una contradicción disfrazada entre "*p*" y "no *p*"[13]. En el caso de (2) es fácil reconocer tal resultado porque, si decir "creo que *p*" es solamente otra alternativa a decir "*p*", afirmar (2) es afirmar (2**) "*p* y no *p*". Sin embargo, en el caso de (1) parece inútil dar solamente este paso ya que podemos usar "No creo que *p*" para describir una situación en la que falta la creencia en *p* o en la que ignoramos si *p*. Así, afirmar (1) no nos lleva directamente a una contradicción entre *p* y no *p* porque "no creo que *p*" no es solamente otra manera de decir "no *p*". Aquí, por lo tanto, la diferencia entre una OM del tipo (1) y del tipo (2) es todavía más importante, porque necesitamos condiciones complementarias para solucionar la primera. Esta es una de las críticas de peso al enfoque wittgensteiniano.

No obstante, hay quien interpreta las observaciones de Wittgenstein de una manera más amplia a la que estamos considerando, y que entiende que hay una doble dirección en la estrategia de mantener que "creo que *p*" es solamente otra manera

12 "Si hubiera un verbo con el significado de 'creer falsamente', no tendría sentido usarlo en la primera persona del presente de indicativo" (PI, parte II, sección X).

13 Linville y Ring (1991: 296) entienden el ejemplo de Wittgenstein "Creo que está lloviendo pero no está" como un absurdo, por consistir en dos enunciados contradictorios acerca del tiempo.

de decir "*p*". Lee (2001: 360-61) entiende que si Wittgenstein dice que "creo que *p*" significa a grandes rasgos lo mismo que "*p*", la primera parte de la conjunción (1), "*p*", también puede entenderse como "creo que *p*". De tal manera, Lee sostiene que es posible hallar el absurdo en la forma de contradicción también en las oraciones del tipo (1).

La maniobra de Lee es interesante pero puede parecer demasiado *ad hoc* para establecer una salida consistente. Aceptar las dos posibilidades de sustitución (o bien "p" por "creo que p", o bien "creo que p" por "p") apela a principios de sustitución que solamente adquieren su sentido cuando interpretamos las palabras de alguien en busca de contradicciones, algo que, en general, no es lo que caracteriza a la interpretación. Es decir, la sustitución que permitiría encontrar una contradicción sólo emerge si ya tuviéramos como objetivo hallar la contradicción. Pero, en general, no se intenta buscar una contradicción en las palabras de alguien.

De cualquier manera, también quiero insistir en la importancia de la contribución wittgensteiniana a que expresiones como "creo que *p*" no vayan separadas de expresiones como "*p*". Según Linville y Ring (1991: 302), que desarrollan un camino derivado del wittgensteiniano, la mayoría de los acercamientos partirían de la suposición de que la paradoja se establece porque estamos tratando una conjunción entre una oración auto-referencial (que dice algo sobre el hablante) y otra oración que tendría un contenido totalmente independiente de aquel. Al darnos cuenta que las dos partes de la conjunción tienen relación entre sí, somos capaces de percibir que sí hay un problema en una OM. En el caso específico de Wittgenstein, pienso que su posición revela la fina intuición que dice que cuando uno se pregunta a sí mismo si cree en *p*, lo que hace es mirar hacia el mundo en lugar de mirar hacia sí mismo.

Por ello, tal solución nos da espacio para que la entendamos como una prueba de que cuando alguien dice algo sobre sí mismo – como "creo que *p*"– lo que tiene en mente es el mundo. En su famosa

cita, Evans (1982: 225), dice que si alguien pregunta "¿Te parece que habrá una tercera guerra mundial?", aquello en lo que pienso para contestar es lo mismo que pensaría si estuviera preguntando "¿habrá una tercera guerra mundial?". Es decir, contesto a tal persona de la misma manera.

S. Shoemaker

La tercera vía para disolver la paradoja está inspirada por Shoemaker[14]. Tal alternativa puede llamarse 'psicologismo sobre la paradoja de Moore' –como sugiere Kriegel (2004: 102)– y afirma que el carácter paradójico de las oraciones mooreanas se hereda de las "creencias mooreanas". Las oraciones mooreanas son absurdas porque expresan creencias mooreanas (Kriegel, 2004: 102).

Lo que sugiere este tipo de solución es que más problemático que afirmar una OM es creer en una. La paradoja no aparece solamente en el ámbito del discurso. De hecho, alguien podría defender que, en la práctica, muchas de las OM podrían incluso ganar sentido. De manera que lo que sí seguiría siendo problemático serían las OM en el nivel de las creencias. En cierta forma, es esto lo que Shoemaker (1995: 75-6) defiende cuando denuncia la poca atención que se presta a la rareza de la idea de que alguien pudiera creer en un contenido proposicional con la forma de una OM. E insiste en que lo que realmente necesitaría explicarse sería por qué una persona no puede creer coherentemente que esté lloviendo y que no cree que está lloviendo.

Según Shoemaker (1995: 76) si una persona creyera en una OM inevitablemente estaría creyendo en una contradicción. Dice:

> Considere la proposición que es la conjunción de esta proposición (Está lloviendo pero no creo que esté lloviendo) y la proposición de que el hablante cree tal proposición, o sea, la proposición expresada por la oración "Está lloviendo y no creo que está lloviendo, y que

14 Kriegel (2004) defiende una variante de tal solución.

> esto sea así es lo que creo". Esto es auto-contradictorio. De este modo, es una característica de los contenidos de las oraciones mooreanas que si uno puede creer de hecho en ellas, el sujeto de tal creencia no podría creer que las tiene sin estar creyendo en una contradicción. (Shoemaker, 1995: 76)

Shoemaker llega a este resultado por medio de algunas condiciones suplementarias, donde la principal es la caracterización de la relación entre creencias de primer y de segundo orden.

De acuerdo con Shoemaker (1995: 77) creer en algo lo compromete a uno a creer que cree en tal cosa, en el sentido de que si tal persona se pregunta si tiene tal creencia, deberá llegar a la conclusión que cree que cree. Es decir:

> Si x cree que *p*; entonces x está comprometida con la creencia de que ella misma cree que *p*.
>
> X está comprometida con la creencia de que *p*, si y solamente si, [...] si x toma en consideración si cree que *p*, entonces vendrá a creer que cree que *p*. (Kriegel, 2004: 106)

De tal modo, la ocurrencia de una creencia mooreana implica en que si el sujeto se pregunta a sí mismo si cree en una proposición mooreana, esto debería llevarle a encontrarse en una contradicción.

Tal acercamiento entiende que la creencia de segundo orden es, de alguna manera, inherente a la de primer orden. Según la lectura de Kriegel (2004: 108), la creencia de segundo orden es una creencia disposicional y sería esto lo que le permitiría a Shoemaker construirla como ya presente, de alguna forma, en la de primer orden.

Sin embargo, la crítica más inmediata a tal enfoque es que no es necesario que alguien se pregunte si tiene o no tiene determinada creencia para que sea un agente racional. En un caso en que una persona tuviera una creencia en una OM y no se preguntara sobre tal creencia, no habría impedimento alguno para que siguiera creyéndola. En este caso, tal persona no tendría en realidad la creencia de que cree en la proposición mooreana, y por lo tanto, no tendría de

hecho (aunque potencialmente sí) creencias contradictorias (Kriegel, 2004: 107).

Shoemaker defiende el principio de que el pensamiento constriñe el habla. En el caso de las OM, según tal posición, realmente no sería posible que alguien las dijera porque no sería posible creerlas, o por lo menos, no sería posible creerlas sin creer en una contradicción. Sin embargo, como Albritton (1995) bien observa, si alguien dice una OM no diríamos "no, no la crees porque no puedes creerla. Es imposible". Sino que diríamos "¿qué?, ¿qué has dicho?", indicando que es otro el problema involucrado en una OM.

Además de tales complicaciones, creo que una posición cómo la de Shoemaker, que soluciona la paradoja por medio de la relación entre creencias de primer y de segundo orden para indicar la contradicción entre creencias, tiene que comprometerse en algún momento con la tesis de la luminosidad. Es decir, en algún momento tiene que decir que el simple hecho de tener una creencia en *p* me pone en una situación en la que tengo una creencia sobre la creencia en *p*. Creo que los casos de fallo de accesibilidad de uno a su propia mente, como los casos de autoengaño, son indicadores de que hay algún problema con tal tesis[15].

A pesar de estos problemas, una posición como la de Shoemaker conlleva también algunas intuiciones importantes y que no deberíamos depreciar. Una de ellas es que el problema con una OM no existe solamente en el acto de afirmarla, no es solo una transgresión del comportamiento. Es también y primariamente una transgresión acerca de lo que se afirma o se cree (Kriegel, 2004: 113-14). El problema no está solamente en que no existe una contradicción en mis palabras sino que en ellas están involucrados más aspectos de la vida mental del sujeto.

15 Es posible argumentar en contra de la tesis de la luminosidad como lo hace Williamson (2000), que ofrece una prueba de reducción al absurdo a partir de las premisas que la compondrían.

Las tres soluciones ofrecen importantes intuiciones que debemos tener en cuenta. Sin embargo, me gustaría ofrecer una cuarta, inspirada tanto por las soluciones que acabamos de ver, como por la manera davidsoniana de comprender el fenómeno de la comunicación. Mi propuesta sugiere que las diferentes soluciones pueden ser más útiles cuando se ven como complementarias que cuando lo hacen como opciones rivales. Y así, lo que voy a intentar es ofrecer una nueva explicación de por qué una OM es absurda.

3. La irracionalidad de una OM

Una breve recapitulación: He sugerido que la primera manera de entender la paradoja expresa la intuición de que un enunciado sobre el mundo nos muestra algo sobre quién lo dijo. Nos muestra, por lo menos, qué tipo de visión tiene determinada persona, desde dónde ve las cosas y con qué se compromete. La segunda solución, he mantenido que no trataba solamente del camino inverso –la idea de que cuando alguien dice algo sobre sí mismo está diciendo algo sobre el mundo–, sino que, además, nos mostraba el carácter externo de la perspectiva de primera persona. Nos mostraba que cuestiones acerca de *p* y acerca de mi creencia en *p*, desde tal perspectiva, se podrían mezclar hasta el punto en que una fuera sencillamente una forma distinta de expresar la otra.

Me parece que ambas intuiciones son bastante sensatas y las voy a mantener en mi propuesta de solución respondiendo a la siguiente tesis:

> T1: La distinción entre oraciones del tipo "p" y del tipo "creo que p" no corresponde a la diferencia entre aquellos enunciados que se refieren al mundo y los que se refieren a la vida psicológica de un sujeto, de manera que los haga enunciados independientes entre sí.

Esta tesis niega tanto que la forma general de las oraciones-tipo

(sea esta "*p*" o "*creo que p*") como su aseveración en actos de habla específicos determine de antemano que su contenido sea o bien acerca del mundo o bien acerca del sujeto. T1 no tiene que comprometerse con una postura más fuerte que redujera un tipo de enunciado al otro, como ocurrió con las dos primeras soluciones expuestas en la sección anterior. Es posible mantener que "*p*" y "creo que *p*" son enunciados distintos. Sin embargo, lo que T1 afirma es que no tiene sentido hablar en términos de qué referencia tiene cada tipo de frase de una manera apriorística; no hay razón para mantener que oraciones con la forma de "*p*" tienen como referencia solamente al mundo y que oraciones con la forma de "creo que *p*" tienen como referencia sólo la vida psicológica de un sujeto.

En este sentido, el papel de T1 es rechazar la presuposición de que una OM sea una conjunción de dos oraciones sobre dos materias distintas, lógicamente independientes entre sí, pero en conflicto. T1, además, no tiene la intención de reducir ninguna de las partes de una OM a la otra.

Con respecto a la tercera solución de la paradoja vista en la sección anterior voy a mantener una segunda tesis:

> T2: lo absurdo de una oración mooreana se extiende tanto al ámbito del discurso, como al ámbito de la creencia.

T2, en realidad, no responde fielmente a la tercera propuesta porque aquella mantenía la tesis de la primacía de la paradoja en el nivel de las creencias sobre el nivel de las aserciones. Pero sí conserva la intuición de que el problema con una OM seguiría existiendo aunque no fuera dicha. La intención de T2 es dar cabida a tal intuición, pero guiada por la idea de que no hay una separación tajante entre el ámbito de la mente y el ámbito del discurso. Así, con referencia a la tercera propuesta, en lugar de mantener la idea de la primacía de la paradoja en el ámbito de las creencias, voy a mantener solamente que la paradoja se extiende a tal esfera.

Voy a desarrollar también una tercera tesis, pero antes intentaré

dar crédito a la plausibilidad de T1 y de T2 en el sentido que he señalado. Para esto voy a utilizar las consideraciones davidsonianas sobre nuestras interacciones lingüísticas, más específicamente usaré su idea de interpretación radical y las tesis involucradas en ella.

En líneas muy breves, la idea davidsoniana de interpretación radical nos invita a aceptar que la atribución de mentalidad a alguien y la atribución de significado a sus palabras ocurren simultáneamente. Es decir, intentar comprender lo que alguien dice y verlo como un agente que tiene una vida mental, no ocurre en momentos aislados ni distintos.

Pensemos en la famosa situación radical, en la que una intérprete se propone hacer un manual de traducción de un lenguaje de una sociedad completamente aislada de cualquier otra del planeta. Tal situación ya presupone que la intérprete sea capaz de reconocer la existencia de un comportamiento intencional y lingüístico en el interpretado. Ella no podría empezar a traducir cualquier sonido emitido por cualquier ser vivo con el que se encontrara. La identificación de la intencionalidad, o sea, la atribución de una mente a nuestro interpretado tiene que ser parte de la misma tarea de comprender sus palabras. Si "Kurt profiere las palabras 'Es regnet' y bajo las condiciones correctas sabemos que está diciendo que está lloviendo", lo hacemos porque al identificar su enunciación como lingüística e intencional, somos capaces de interpretar sus palabras: podemos decir qué significan, en ese momento, sus palabras (Davidson, 1973: 125).

La propia actividad de describir comportamientos lingüísticos es ya una actividad de interpretación en la que, no solamente atribuimos determinados significados a las palabras de alguien, sino que también atribuimos estados mentales, como creencias y deseos. Traducir 'Gavagai' sólo puede tener sentido como actividad si el intérprete entiende que tal sonido lo ha emitido un ser que se relaciona de determinada manera con el mundo, muy similar a la suya, y que quiere decir algo cuando dice 'Gavagai'.

Es necesario notar que la visión de Davidson puede, y de hecho se extiende, desde la situación radical hasta nuestras situaciones más cotidianas. Según propone Davidson (1973: 125) toda comprensión del discurso del otro involucra una interpretación radical que podría entenderse, para hablantes de un mismo lenguaje, en términos de la pregunta: ¿cómo se puede determinar que el lenguaje es el mismo?

Si tal enfoque es plausible, parece que tanto T1 como T2 ganan fuerza. Alguien que niega T1 parece estar de acuerdo con que es posible oír "*p*" y no tener en cuenta nada de lo que representa que "*p*" sea dicho por alguien. En el caso de T2, alguien que la negara parece tener que comprometerse con la posibilidad de que hubiera algo en el ámbito de las creencias que no fuera accesible a nadie, o sea, que existieran cosas en la mente sin que existiera nada en el habla. Sin embargo, como he dicho, creo que tal contexto solamente es un estímulo para que uno acepte T1 y T2. No hay una consecuencia directa entre las posiciones davidsonianas y tales tesis, principalmente porque estamos hablando, en el caso de una OM, de oraciones particulares, mientras que Davidson defiende tesis generales para el lenguaje y la interpretación. Si tal enfoque es correcto, no hemos demostrado que "*p*" no refiera sólo al mundo ni "no creo que *p*" sólo al hablante, pero creo que nos anima a defenderlo. Y nos motiva a ser más holistas en ese respecto. Parece ser posible decir que tanto "*p*" como "creo que *p*" puedan hacer explícitos, a la vez, datos sobre el mundo y datos sobre el agente en el mundo.

Defender la plausibilidad de T1 y T2, sin embargo, tampoco explica qué equívoco hay en una OM. Para esto, antes de localizar donde se encuentra lo absurdo de una OM, expongo mi tercera tesis:

> T3: Un enunciado es absurdo cuando no responde a los principios de la racionalidad[16].

16 Los principios de la racionalidad a que T3 se refiere son más amplios que la noción de coherencia y consistencia lógica, como se mostrará más adelante.

T3 llama la atención sobre otro presupuesto de la discusión tradicional sobre la paradoja de Moore y lo niega: que un enunciado sólo sería absurdo en presencia de una contradicción. Parece claro que un enunciado puede ser absurdo debido a muchas razones. No obstante, la paradoja se entiende tradicionalmente sólo en términos de la falta de una contradicción: "sabemos que una OM es absurda aunque no haya una contradicción presente en la oración en sí misma"[17]. La intención de T3, especialmente junto a T1, es dar cabida a una solución sin que tengamos que pasar por la reducción de una parte de la OM a otra para encontrar la oración contradictoria. Lo que T3 propone es que, resolver la paradoja no es indicar donde está la contradicción en la OM, sino explicar por qué una OM no es una aserción racional.

De vuelta al mismo panorama davidsoniano sobre la interpretación radical, otra tesis involucrada allí afirma que la actividad de interpretación solamente tiene éxito si se atribuye al interpretado un buen grado de consistencia y coherencia en sus comportamientos lingüísticos, además de un buen grado de acierto en sus enunciados. Esto hace parte de lo que Davidson llama 'principio de caridad' que no es solamente la indicación de una estrategia para que interpretemos a alguien, sino la única manera que tenemos para interpretarnos mutuamente, sea esta una situación radical o cotidiana. Podemos entender, por lo tanto, que enunciados tomados como significativos e intencionales sólo pueden interpretarse si entendemos las palabras del interpretado sobre un fondo de coherencia y consistencia. Defenderé además de esto que, es necesario que supongamos que existe una unidad personal en nuestro interpretado. Llamaré a todo este conjunto de suposiciones que hacemos al interpretar a alguien 'la suposición de la racionalidad' de nuestro interpretado.

Propongo que nos acordemos de la situación imaginaria del

17 Williams (2006), por ejemplo, define la paradoja en estos términos.

inicio del texto, en que mi amiga dice "está lloviendo pero no creo que esté lloviendo". Como he dicho, estamos las dos frente a una ventana por la cual se puede ver claramente que llueve. Yo podría, por un lado, intentar reinterpretar sus palabras en otros términos, como, por ejemplo, una metáfora o como si quisiera decir otra cosa como "!no creo que esté lloviendo! Se han estropeado mis planes de ir a la playa"[18]. Pero si no lo hago, tendría que ver que mi amiga claramente no tiene un comportamiento racional. Como he sugerido, sé que se equivoca cuando dice que no cree que llueva, pero me deja sin espacio para convencerla de que está equivocada porque ha dicho al principio "llueve". ¿Se ha equivocado o sabe que llueve? Parece que para entenderla, tendría que dividirla en dos personas, una que afirma la primera parte de su enunciado y otra que afirma la segunda parte. Sin embargo, no podría verla de tal manera, con tal división, sin pensar en ese momento que está siendo irracional.

En su trabajo de 1982, Davidson presenta un enfoque bastante interesante sobre la irracionalidad, y la toma como siendo un fenómeno que ocurre "dentro de la casa de la razón" y no ya fuera de la esfera de lo racional. De tal forma, la irracionalidad no se establecería para aquellas acciones de los demás que, bajo nuestro juicio, son poco razonables, sino para las creencias y acciones que resulten de un fallo dentro del ámbito mental de la propia persona. Según la definición que ofrece Davidson (1982: 179), serían casos de irracionalidad aquellos en los que hubiera una causa mental que

18 En este caso, si mi amiga quisiera decir esto, su enunciado no sería un caso genuino de OM. De hecho, existen dichos populares que tienen exactamente la misma forma que una OM, pero no lo son, porque son capaces de transmitir un sentido inteligible. Por ejemplo, el dicho "no creo en políticos honestos, pero haberlos haylos", puede usarse perfectamente para mostrar la perplejidad frente a la existencia de tantos políticos corruptos. Otro uso común podría ser la oración "São Paulo es una ciudad muy desagradable, pero creo que no lo es", para expresar la idea de que aunque la mayoría de la gente la considere así, a mí no me parece que así sea. Es interesante que tales oraciones se parecen mucho a una OM, pero al utilizarlas expresando un sentido inteligible, no pueden entenderse como tal.

no sería una razón para lo que causan, y es solamente en este sentido en el que podríamos decir que una persona es irracional cuando no está abierta a razones.

Pienso que este enfoque es bastante interesante para tratar de la irracionalidad porque por un lado, es más general que aquellos que entienden la irracionalidad como fallos en las reglas predeterminadas del razonamiento. Y por otro lado, la trata sin transformarla en otro fenómeno, es decir, consigue tratarla sin tomarla como un caso inexistente[19].

Al definir casos de irracionalidad como casos de estados mentales que tienen como causa otros estados mentales que, sin embargo, no sirven de razones para los primeros, Davidson sostiene que si queremos explicar de hecho la irracionalidad, deberíamos asumir que la mente se divide en estructuras cuasi-independientes (Davidson, 1982: 181). Es solamente de ese modo como un estado mental podría causar otro sin jugar el papel de razón para él. Sostener tal idea no significa tener que enumerar, ni nombrar las diferentes partes de la mente. Tampoco necesitamos hablar de divisiones tajantes entre tales partes. Podemos mantenernos en un nivel más abstracto que este. Davidson señala incluso que términos como 'partes de la mente', o 'partición de la mente' son términos engañosos si con ellos entendemos que lo que está en una región no puede estar en la otra.

Frente a tal acercamiento, he sugerido que una OM auténtica debería entenderse como un caso de irracionalidad porque hace evidente la división interna de la persona: en el ejemplo propuesto,

19 Varios enfoques sobre la irracionalidad, según Davidson (1982), acaban por diluir el propio fenómeno. El principio de Platón, por ejemplo, que se basa en la doctrina de la racionalidad pura, diluiría la irracionalidad, transformándola en un caso más de ignorancia. En el extremo opuesto, según el principio de Medea, un acto irracional sería fruto de fuerzas ajenas a la persona. Según tal posición en un caso de acracia la persona actuaría en contra de su mejor juicio porque alguna fuerza ajena superaría a su voluntad, de lo que se sigue que los actos acráticos no serían intencionales.

mi amiga se encuentra en el mundo en parte considerando que llueve y en parte considerando que no llueve[20]. Esto vuelve a mi amiga inconsistente, pero no a sus palabras.

En las condiciones de una interpretación radical, he mantenido que si queremos interpretar a alguien, debemos suponer la racionalidad de nuestro interpretado. Y suponer esto significa atribuir un fondo de verdad y de coherencia a sus palabras, y, además, la atribución de una unidad a su mente. En un cierto sentido, el "principio de caridad" davidsoniano es aquí ampliado porque la mera coherencia (del sistema de creencias, así como entre lo que se cree y lo que se dice), o incluso, la presunción de verdad en las palabras de alguien, no da plena cuenta de la racionalidad involucrada en la unidad personal. La explicación davidsoniana de la irracionalidad

20 Moran (2001: 69), por ejemplo, sugiere que casos de acracia o autoengaño, que son tradicionalmente tomados como fallos de la racionalidad, pueden entenderse en términos de la paradoja de Moore, justamente porque aparecen en situaciones de verdadera división con respecto a cómo uno se entiende a sí mismo: la división entre una actitud que tengo y otra que me atribuyo. Ofrece el siguiente paralelo: "sé que lo que siento con respecto a él deberían ser celos ("creo que está lloviendo"), aunque no haya nada de lo que tener celos" ("pero no está lloviendo"). Según lo expone, aquí tampoco tenemos una contradicción, porque tal enunciado representa un estado de cosas posible, pero sabemos que hay una gran tensión en tal caso.

Moran tiene un enfoque muy interesante sobre cómo tenemos acceso a nosotros mismos y una de sus tesis es que lo hacemos de dos modos: desde la perspectiva de primera persona y desde la perspectiva de tercera. Esto no significa que tengamos dos tipos de objetos distintos o partes específicas de lo mental accesibles a cada uno de los modos. Los dos modos de acceso son solamente distintas maneras de acceder a nosotros que nos posibilitan retractarnos o corregir nuestras propias creencias. Podría tener acceso a mí misma tanto por medio de uno como de otro, y corregirme si fuera el caso, justamente por la posibilidad de contrastar los dos puntos de vista sobre mí misma y mis comportamientos (Bensusan & Pinedo, 2007). Comparado con la paradoja de Moore, un caso de autoengaño, por ejemplo, sería tan poco racional como un enunciado de una OM si el sujeto siguiera manteniendo dos puntos de vista contrapuestos sobre sí mismo. La posibilidad de que tengamos dos vías de acceso a nosotros mismos y que lleguemos a veces a resultados contrarios no nos da derecho a mantenerlos simultáneamente.

nos empuja a tomar la mente como compartimentada, pero esto no significa que tengamos que considerarla así en el momento de la interpretación. Al contrario, no podemos hacerlo. Según el propio Davidson, no hay dudas de que el principio de caridad exigido en una actividad de interpretación se opone a la idea de partición de la mente (Davidson, 1982: 184). Es decir, la irracionalidad se explica en términos de la posibilidad de la fragmentación de la mente, pero si de hecho no tomamos a nuestro interpretado como alguien irracional, no tenemos que considerarlo como fragmentado.

El problema con una OM podría formularse de la siguiente manera: en una situación de interpretación radical, sencillamente no podríamos empezar atribuyendo una OM a nuestro interpretado. Precisamente porque una OM es la expresión de un comportamiento irracional y, como vimos, tenemos motivos para aceptar que interpretar solamente es posible si asumimos que estamos frente a un ser racional. Está claro que tenemos comportamientos irracionales y es imprescindible que mantengamos un espacio en nuestras teorías para que quepan allí. Pero lo importante es que esto no puede ser la norma. En el caso de la lluvia, si estuviéramos tratando de un lenguaje totalmente desconocido, entonces no podríamos empezar atribuyendo algo como una OM a las palabras de mi amiga.

Estamos obligados a tomar las palabras de alguien como teniendo que ver con su posición en el mundo y como manifestación de cómo esta persona en particular se relaciona con él. Pero además, buscamos que tal persona no se encuentre dividida en diversas partes que se relacionen de manera independiente unas de las otras, con los demás y con el mundo. Por esto, T3 tiene todavía más plausibilidad. Una OM es problemática, no porque haya una contradicción en enunciarla, sino porque nuestra idea de racionalidad supone la unidad personal. No se puede atribuir una OM bajo los principios de la racionalidad porque además de ir contra nuestras intuiciones lógicas también choca con nuestra concepción de nosotros mismos como personas. Y una de las ideas esenciales es que hay una unidad en

cada uno de nosotros. Una idea, que a su vez, es totalmente distinta de la noción de accesibilidad total a todos nuestros estados mentales, y que tampoco es prueba de que no tengamos comportamientos contradictorios o tensiones dentro de nosotros mismos.

De forma que, creo que T1, T2 y T3 establecen un contexto desde donde podemos entender la peculiaridad de una OM con referencia a cómo estamos comprometidos con la idea de que actuamos lingüísticamente como seres racionales. Y en este caso, actuar lingüísticamente no puede suponer que las esferas del habla y de las creencias puedan estar separadas. Ni puede suponer que mantengo una posición neutra cuando hablo sobre mí o sobre el mundo; en ambas circunstancias aparecen rasgos sobre lo que pienso y sobre cómo veo el mundo. Si digo algo que solamente podría ser inteligible si me tomaran por dos personas, entonces es bastante probable que esté diciendo una OM.

Referencias Bibliográficas

Albritton, R. 1995, "Comments on 'Moore's Paradox and Self-knowledge'", *Philosophical Studies* 77 (2/3), pp. 229-39.

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

Bensusan, H. & Pinedo, M. 2007, "When my Own Beliefs are not First-personal Enough", *Theoria* 58, pp. 35-41.

Black, M. 1952, "Saying and Disbelieving", *Analysis* 13, pp. 25-33.

Davidson, D. 1973, "Radical Interpretation", en D. Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 125-39.

Davidson, D. 1982, "Paradoxes of Irrationality", en D. Davidson 2004, *Problems of Rationality*, Oxford, Clarendon Press, 169-88.

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

Heal, J. 1994, "Moore's Paradox: A Wittgensteinian Approach", *Mind* 103, pp. 5-24.

Hintikka, J. 1962, *Saber y Creer* (traducción J. J. Acero), Madrid, Editorial

Tecnos, 1979.
Kriegel, U. 2004, "Moore's Paradox and the Structure of Conscious Belief", *Erkenntnis* 61, pp. 99-121.
Lee, B. D. 2001, "Moore's Paradox and Self-Ascribed Belief", *Erkenntnis* 55, pp. 359-70.
Linville, L. & Ring, M. 1991, "Moore's Paradox Revisited", *Synthese* 87, pp. 295-309.
Moore, G. E. 1942, "A Reply to My Critics" en P. Schilpp (ed.), *The Philosophy of G. E. Moore*, La Salle, Open Court, pp. 535-677.
Moore, G. E. 1944, "Russell's 'Theory of Descriptions'" en P. Schilpp (ed.), *The Philosophy of Bertrand Russell*, Evanston, Northwestern University, pp. 175-226.
Moran, R. 2001, *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton, Princeton University Press.
Shoemaker, S. 1995, "Moore's Paradox and Self-Knowledge", en S. Shoemaker 1996, *The First-person Perspective and Other Essays,* Cambridge, Cambridge University Press, pp. 74-93.
Williams, J. N. 1979, "Moore's Paradox – One or Two?", *Analysis* 39, pp. 141-42.
Williams, J. N. 2006, "Wittgenstein, Moorean Absurdity and its Disappearance from Speech", *Synthese* 149, pp. 225-54.
Williamson, T. 2000, *Knowledge and its Limits*, Oxford, Oxford University Press.
Wittgenstein, L. 1953, *Investigaciones Filosóficas* (traducción A. García Suárez & U. Moulines), Barcelona, Crítica, 1988.

Translation of Chapter:

5 Interpreting Moore's Paradox: the Irrationality of a Moorean Sentence[1]

Published in Spanish in *Theoria*, vol. 23/2, num. 62, pp. 145-161, 2008.

ABSTRACT
This paper offers an interpretation of Moore's Paradox that emphasizes its relevance for our understanding of rationality and linguistic interpretation. The sentences that originate the paradox do not need to be thought of in terms of the absence of a contradiction, but in terms of absence of rationality, where rationality is understood as a broader notion than coherence and logical consistency. This is defended through three theses, two of which stem from the dominant (but insufficient) approaches to the paradox: Moore's, Wittgenstein's and Shoemaker's.

KEYWORDS: Moore's paradox, rationality, radical interpretation, personal unity, externalism and first person perspective.

Introduction

Suppose that a friend and I are in front of a window, from where

1 Previous versions of this work have been presented and discussed at the 5th Works in Progress Workshop, in March of 2007 at the University of Granada, and at the 5th SEFA congress, in September of 2007 in Barcelona. I'm grateful to Manuel de Pinedo García, Neftalí Villanueva Fernández, Mar Muriana López, Miguel Ángel Pérez and Theoria's two anonymous referees.

we see that it is raining. She says: "It is raining". I listen to her and understand that she wants to say something about the weather; she wants to draw my attention to an event that we are both observing. But then, she states "but I don't believe it is raining". What should I answer to her? On the one hand, I see that it is raining and I know that she is wrong in claiming that she does not believe it is raining. On the other hand, I just cannot give her any reason to make her change her mind because the first thing she said was "it is raining". It seems that the only thing I could do would be to insist "Excuse me?". Despite this dilemma, I am forced to accept that my friend has uttered grammatically acceptable sentences.

The discussion of Moore's paradox is placed in contexts that are similar to the above one, where a statement that in principle does not have any formal problem turns out to be an absurdity. Traditionally, the paradox is characterized in terms of a consensus about the absurd character of such utterances whereas there is no contradiction in the sentence itself. Because of this, the debate has usually been directed to the search for what could be a disguised contradiction within such sentences.

In this work, I will offer an interpretation of the paradox that emphasizes its relevance for our understanding of rationality and linguistic interpretation. I will defend that we do not need to focus on the absence of a contradiction in such paradoxical sentences. The paradox is better understood in terms of absence of rationality. My position involves the defense of three theses; two of them stem from the dominant (but, in my opinion, incomplete) approaches to the paradox: Moore's, Wittgenstein's and Shoemaker's[2]. I will argue that Moore's paradox is not a self-contained puzzle. The way

2 I am grateful to the anonymous referee who helped me to see the importance of Hintikka's (1962) solution. Nevertheless, I'll discuss such a solution along several passages of this paper instead of considering it as the fourth dominant approach to the paradox. As I will point out, Hintikka's ideas share common aspects with some of the solutions that will be studied here, including my own.

to solve it depends on issues that are central for our understanding of self-attribution of knowledge, mental externalism and linguistic interpretation.

First, I will introduce Moore's paradox. Next, I will present the three dominant approaches to the paradox and my solution.

1. What it is paradoxical about "Moorean sentences"

Consider the following sentences

i. *p*
ii. I believe that *p*
iii. I do not believe that *p*
iv. I believe that not-*p*

(i) could occur jointly with any of the cases expressed by the other sentences: *p* could be the case and I could believe that *p*; *p* could be the case though I could ignore it; and finally, *p* could be the case and I could mistakenly believe the negation of *p*. Despite the plausibility of these situations, asserting any of the following sentences:

1. *p* but I do not believe that *p* (i & iii)

and

2. *p* but I believe that not-*p* (i & iv)

generate very strange situations.

It is here where "Moore's paradox" arises. Henceforth, sentences of the type (1) and (2) will be called 'Moorean sentences' (MS). The world can easily accept something being the case without my believing it, but we must refuse someone's words if she states a fact while at the same time she states her believing its negation. The same occurs when someone asserts both a fact and the absence of her belief in such a fact. Traditionally speaking, there is no contradiction either between (i) and (iii) or between (i) and (iv), but even so, when such conjunctions are asserted by the first-person of the discourse, they become unacceptable; they are absurd.

According to Williams (2006: 226) "Moore was careful to

distinguish absurdity from paradox". In the case of MS "what is absurd is for me to assert such sentences. What is paradoxical is that this absurdity persists in the absence of semantic contradiction in my words themselves" (Williams, 2006: 227).

Along the text, I will argue that such sentences are in fact unacceptable, but because they do not respond to rational behavior. They are not mere cases of logical misunderstandings. Nevertheless, by now, I will continue by specifying the sort of sentences we are dealing with. It is worth noting the peculiarity of (1) and (2) when compared to some other sentences that are not MS despite being very similar to them, such as the following ones:

3. *p* but she does not believe that *p*

4. *p* but I believed that not-*p*

5. *p* but I know that not-*p*

There isn't anything contradictory or paradoxical in stating either (3) or (4). Both sentences could be perfectly coherent. The paradoxical aspect of "*p* but I do not believe that *p*" in contrast with "*p* but she does not believe that *p*" (or even "*p* but I believed that not-*p*") invites us to recognize the difference between the first- and third-person points of view. Although (1) and (3) have the same logical form ($p \,\&\, \sim B_S p$), the difference between one being paradoxical and the other not, lies in who S is; who is the subject who believes *p*[3]. By asserting (1), I am speaking about myself whereas by asserting (3) I am speaking about another person. And there are plenty of times when we perceive an event that contrasts with what others believe.

The fourth sentence is also familiar. There are several circumstances when we realize that something is the case although

3 Hintikka (1962: 66) emphasizes this difference and recognizes that the first-person pronoun 'I' has an important role here. However, he suggests that not only it would be preferable to explain the paradox without adding a criterion that makes reference to it, but also that this sort of explanation "in terms of the peculiarities of first-person pronouns is in principle misguided" (Hintikka, 1962: 66).

we have being mistakenly believing its opposite. In this case, the sentence is stated by me, about myself, but also from a third-person perspective. I have to look and understand myself in the past, holding a belief that I do not have anymore. In order to do so, it seems to be necessary to distance myself from my usual perspective, adopting a third-person one.[4] In the fourth example, one gets rid of being in an absurd situation by locating one's belief in a different time.

The fifth sentence is very different from the previous ones because we can find in it a clear contradiction between *p* and not-*p*. And that is due to the facticity of knowledge. In (5), I am mistaken because I am affirming a contradiction. It is not a paradoxical sentence because we do not have problems in explaining why the sentence is unacceptable. If someone knows that *p*, it is implied that *p* is the case. There is no mystery here.[5]

It is possible to establish several other relations between Moorean sentences and sentences which only seem to be Moorean. Nevertheless, the previous cases should be sufficient to mark some specificities of a MS and to motivate the idea that a real Moorean sentence is problematic. It is time to ask ourselves: What is the problem with a MS? In the following section I outline some different answers to this question.

4 Someone could argue against the idea that assuming a third-person perspective helps one getting oneself rid of Moore's paradox. After all, one could assume such a perspective also to state one's own present mental states, for example, with the help of a psychologist. In that case, one could also conclude something with the form of a MS. Moran (2001) emphasizes such cases, when one's first-person perspective conflicts with a third-person one. I totally agree with the possibility of such cases. Nevertheless, the peculiarity of (4) lies in the fact that the sentence in question shows the different kind of access I have to myself when I say that I do not have a belief that I had in the past.

5 T. Williamson (2000) defends that sentences of the type (5) can be understood as paradoxical Moorean cases.

2. Solving the paradox

Solving the paradox amounts to discovering where the absurdity of a MS resides. Some have suggested that in order to solve Moore's paradox we must find the contradiction-like phenomenon somewhere else than the sentence in its original form. There are several proposals concerning such an enterprise, but at least three of them are especially relevant: the one inspired by Moore's remarks, another one developed by Wittgenstein, and the third one by Shoemaker. I will analyze them in order to highlight the main insight behind each solution.

G. E. Moore

The philosophical relevance of Moore's paradox (as well as its baptism) is due to Wittgenstein (1953). According to Baldwin (1996), Moore presented his paradox in two late works and in neither of them the paradox was the main theme. He used the sentences that originated the paradox only "to illustrate the distinction between what someone says, or asserts, and what he implies by saying what he does" (Baldwin, 1996: 226).

It is possible, however, to find a solution inspired by Moore in such works. With his famous example: "I went to the pictures last Tuesday, but I don't believe that I did" (Moore, 1942: 543), Moore holds that my statement that I went to the pictures implies that I believe that I went. And because of this, he maintains that stating such a sentence would be absurd.

One possible interpretation of the idea that "my statement that I went to the pictures implies that I believe that I went" would be to maintain that (i) logically implies (ii)[6]. Under this interpretation,

6 I believe that such an interpretation is not fair to Moore's position itself. However, the interpretation that does offer a fair reading of his positions doesn't reach a satisfactory solution insofar as it seeks to find a contradiction and fail to do so. I

if someone asserts (i), we could infer (ii), and this would allow us to see the tension we were looking for. In the case of (1) *–p* but I do not believe that *p*– if we follow such a logical sense of 'implication', we could infer (1*) "I believe that *p* but I do not believe that *p*"; whereas in the case of (2) *–p* but I believe that not-*p*– we would have (2*) "I believe that *p* but I believe that not-*p*".

The difference between sentence (1) and sentence (2) [as well as between their resulting (1*) and (2*)] could be, in principle, considered as being too weak to mark a real difference, since we normally use the sentence (iii) "I do not believe that *p*" as meaning the same as (iv) "I believe that not-*p*". However, Williams (1979) suggests that such a difference must be taken into consideration because in the case of Moore's Paradox, this difference marks two problems where, initially, there was only one.

According to Williams:

> [T]he absurdity of (1) differs from (2). For normally, it is absurd for A to assert [1] because what is conjointly expressed and asserted, i.e. a belief that *p* and a lack of belief that *p*, is *logically impossible*. The absurdity in (2) is of a different kind. For normally, it is absurd for A to assert [2], not because what is conjointly expressed and asserted, i.e. a belief that *p* and a belief that it is not the case that *p*, is logically impossible, but because it is *inconsistent*. (Williams, 1979: 142)

I will keep considering (1) and (2) as being different cases, but I will not commit myself to Williams' terminology because it only makes sense if we have already accepted this first interpretation; an interpretation that, however, doesn't seem to be faithful to Moore's position itself. Although Moore doesn't develop a systematic solution to the paradox, one can understand from his writings that the sense of the implication which he refers to is not the logical sense, but instead an informal and ordinary notion[7]. In one of his works,

---

will deal later on with this issue. I'm considering the above interpretation because it is a reading adopted by other philosophers, such as Williams (1979).

7 There are strong reasons to understand that when Moore talks about implication,

Moore (1944: 204 *apud* Black, 1952: 26) maintains that if we say, for example, "he has not gone out" we imply that we don't believe that he has gone out, although this hasn't been asserted nor follows from anything that was asserted.[8] According to Moore, we imply such a thing because we lead our listeners to assume it. After all, "people don't make, in general, a positive assertion unless they don't believe its opposite is true" (Moore, 1944: 204). Because of this, the first interpretation that promised a quick solution to the paradox fails to do so. In part, because this is not Moore's position, but mainly because it is not so simple to support a logical implication between (i) and (ii). If someone can suppose (ii) when someone asserts (i), this is not the same as maintaining that (i) implies (ii). And Moore is not naive about it.

We are left, then, with the other interpretation of Moore's position, where 'implication' (between my assertion that *p* and my believing *p*) is understood as a weaker relation than the one involved in logical implication. Bearing this in mind, Linville & Ring (1991: 295-296) point out that MS are contradictory only if

---

he doesn't refer to logical implication, but to epistemic or pragmatic implication [this is something that, for example, Hintikka (1962) insists on in his interpretation of the paradox]. Max Black (1952), not using yet terms as precise as the ones applied by contemporary pragmatics, offers an interpretation about Moore's usage of the notion of inference which can result enlightening: "From these remarks of Moore, we can derive the following explanation of the way in which he is here using the word 'imply': Suppose (i) a speaker is using an expression, E, (ii) people do not generally use E unless some related proposition pE is true, (iii) people hearing the speaker use the expression E will generally believe pE to be true, and, finally (iv) the speaker knows all this –then if these four conditions are met, the speaker's words may be said to imply pE" (Black, 1952: 26).

8 In this context, we could make use of pragmatic notions such as statement or utterance understood as a speech act that is subject to certain norms, such as the assumption that someone who states (utters) something believes what he states (utters). But, again, it is important to point out that reading Moore in terms of contemporary pragmatic categories is one among several possible interpretations, since he did not have such resources for his works.

we understand implication according to the second sense (i.e., the logical sense). Moreover, once the paradox is understood in terms of finding a contradiction, this second interpretation of Moore becomes unsatisfactory.[9]

However, it is possible to find virtues in Moore's contribution, although he doesn't fully solve the problem. An important idea that is motivated by his suggestion is the idea that if someone asserts *p*, such a person would be, most of the time, ready to back up such an assertion, in the sense of committing herself to *p*. I believe that such

9 Another possible solution would be to follow Hintikka (1962), who appeals to the principles of believing and of knowing to explain the paradoxical character of a MS. Hintikka recognizes the importance of Moore's point about the existence of a sort of presupposition in communication about the fact that people believe what they say. However, he goes further than that. According to Hintikka "Moore bases his explanation of the oddity of [MS] on the fact that in the great majority of cases we believe what we say" (Hintikka, 1962: 95). But he later emphasizes: "I do not have to assume that, whenever one makes a statement (say, utters q) under normal conditions and in a normal tone of voice, there is a presumption that one believes that q is true. It suffices for my purposes to assume that, whenever one utters q, there is a presumption that one conceivably can believe what one says (in the sense that q must not be obviously indefensible)" (Hintikka, 1962: 98). Moreover, he says: "If I am right, whoever utters [a MS] says something it is impossible for him to believe in the long run […] And this impossibility is a consequence of the logical properties of [MS]" (Hintikka, 1962: 96). Although Hintikka seems to save the Moorean intuition about the problem, he goes further when he indicates that the peculiarity of a MS is understandable through the concept of doxastic indefensibility (Hintikka, 1962: 92). In this sense, such a solution would share important traits with Shoemaker's account, which defends that the real problem with a MS would be to believe one instance of it. I will discuss this issue later on. But, again, Hintikka could not be reduced to such a position because he accounts for the following "I do not think that only such forms of words make sense as can be uttered and believed (or known) by my utterer. The fact rather seems to be that forms of words which do not meet his requirement are useless for most of the purposes which our language is expected to serve" (Hintikka, 1962: 98). I thank the comments of an anonymous referee who indicates that the account I defend in this paper in terms of absence of rationality could share some aspects with Hintikka's account, who takes the "principles of believing and knowing" as the explication of its absurd nature.

an account allows us to preserve the intuition that when one asserts *p*, what is said deals not only with the world, but also with oneself, at least, in the sense of making explicit one's point of view.

When I presented the paradox, the fact that the conjunction involved a statement allegedly about the world and another one allegedly about the subject's belief, was what prevented us from finding a contradiction. Although Moore doesn't find one either, I believe his effort is one of bringing closer the subject matters of the two statements. In the case of Moore, he suggests that assertions about the world are not detached from the existence of someone who asserts them. And if someone asserts them, we understand that she is committed to what she says.

L. Wittgenstein

The second solution is inspired by Wittgenstein[10]. Such a solution can be understood as the inverse strategy to the first one. Wittgenstein suggests that saying "I believe that *p*" is just another way of saying "*p*". In his own words: "[T]he statement 'I believe it's going to rain' has a meaning like, that is to say a use like, 'It's going to rain', but the meaning of 'I believed then that it was going to rain', is not like that of 'It did rain then'" (PI, part II, section X).

Heal (1994: 20) interprets this proposal as the idea that a person who learns a language is trained to employ, in some occasions, "I believe that *p*" as a substitute for the plain assertion "*p*". The alleged reference to the person's belief that *p* would not add anything to what is said by "*p*". In this sense, more than inverting the direction of Moore's solution, what characterizes this second option is that it explains where the contradiction is without referring to what the person implies with his words. The verb 'to believe', when used by

10 Heal's (1994), Lee's (2001) and Linville & Ring's (1991) accounts are all examples of variations of the Wittgensteinian solution.

the first-person, would be better understood as a redundant term[11].

If we apply this reasoning to the paradox, we will find that a MS is just a disguised contradiction between "*p*" and "not-*p*"[12]. In case (2) it is easy to reach this result, because if saying "I believe that *p*" is just another way of saying "*p*", to assert (2) is to assert (2**) "*p* and not-*p*". However, in the first case, it is useless to take just this step. For we can use "I do not believe that *p*" to describe a situation where we lack the belief that *p* or where we ignore *p*. So, asserting (1) does not lead us straight to a contradiction between *p* and not-*p* because "I do not believe that *p*" is not just an alternative way of saying "not-*p*". Here, the differences between MS of type (1) and type (2) are even more crucial, because we need further conditions to solve the paradox in the first case. And this is an important criticism to the Wittgensteinian approach.

There are, however, broader interpretations of Wittgenstein's solution –such as the one developed by Lee (2001)– that accommodate the idea that not only "I believe that *p*" is just another way of saying "*p*", but also the other way round: "As Wittgenstein says, if 'I believe that P' means roughly the same as 'P', the first conjunct of [1: P] can be taken as the same assertion as 'I believe that P', and so we can explain why an [1]-type Moorean sentence is also absurd" (Lee, 2001: 360-61).

Lee's (2001) strategy is interesting but it can seem too *ad hoc* for establishing a consistent solution. The acceptance of two possibilities of substitution ("*p*" by "I believe that *p*" or "I believe that *p*" by "*p*") appeals to principles of substitution that make sense only if we already interpret someone's words in search of a contradiction; this is something that in general doesn't characterize interpretation.

11 Wittgenstein says: "If there were a verb meaning 'to believe falsely', it would not have any significant first person present indicative" (PI, part II, section x).

12 Linville and Ring understand that Wittgenstein's example "I believe it is raining but it's not" is absurd because it consists of two contradictory assertions about the weather" (Linville & Ring, 1991: 296).

That is, the substitution that would allow finding a contradiction only emerges if we have previously the aim of finding it. However, in general, one doesn't search for a contradiction in another's words while interpreting him.

I also want to insist on the importance of the Wittgensteinian contribution, which holds that expressions like "I believe that *p*" are not detached from expressions like "*p*". According to Linville and Ring (1991: 302) –who defend a variation of the Wittgensteinian solution– most approaches to Moore's paradox depart from the presumption that the paradox appears because we are dealing with a conjunction between a self-referential sentence (that says something about the speaker) and another sentence that has a totally independent subject matter. In noticing that both parts of the conjunction are related to each other, we are able to understand the problem with a MS. In the specific case of Wittgenstein's solution, it can accommodate a very subtle insight: when I ask myself whether I believe *p*, what I do is not to look inside myself, but rather look at the world.

This kind of solution points to the fact that when someone says something about oneself –such as "I believe that *p*"– what one has in mind is the world. In the famous quotation of Evans, he says "(…) 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?'" (Evans 1982: 225).

S. Shoemaker

The third way of solving the paradox is proposed by Shoemaker[13]. This alternative can be called 'psychologism about Moore's paradox' (Kriegel, 2004: 102) and it defends the idea that the paradoxical

13 Kriegel (2004) defends a variation of Shoemaker's solution.

character of Moorean sentences is inherited from Moorean beliefs. In that sense: "Moorean assertions are absurd, when they are, because they express Moorean beliefs" (Kriegel, 2004: 102).

The main idea behind this kind of solution is that believing a MS is more problematic than stating one. The paradox arises not only for discourse. In fact, someone could defend that in practice many MS could even make sense. In such circumstances, the problem would remain at the level of beliefs. Shoemaker emphasizes his perplexity with the little attention that is given to the peculiar idea that someone could believe the propositional content of a MS. He insists that what "really needs to be explained is why someone cannot coherently *believe* that it is raining and that she doesn't believe that it is" (Shoemaker, 1995: 75-76).

According to Shoemaker (1995: 76), if someone believes a MS, she would, in fact, believe a contradiction. He explains:

> [C]onsider the proposition that is the conjunction of this proposition [It is raining, but I don't believe that it is raining] and the proposition that the speaker believes this proposition, i.e, that expressed by the sentence "It is raining and I don't believe that it is raining, and that this is so is something I believe". That is self-contradictory. So it is a feature of the contents of Moore paradoxical sentences that if they can be believed at all, the subject of such a belief could not believe that she had it without believing a contradiction. (Shoemaker, 1995: 76)

Shoemaker reaches this result through the consideration of some extra conditions, such as the characterization of the relation between first and second-order beliefs. According to Shoemaker (1995: 77), believing something commits one to believing that one believes it. If such a person asks herself whether she has this belief, one must conclude that she believes that she believes it. That is:

> If x believes that *p*; then x is committed to the belief that she herself believes that *p*.
>
> x is committed to a belief that *p* iff, if (in circumstances C, yet to

> be specified) x considers whether she believes that *p*, then she will come to believe that she believes that *p*. (Kriegel, 2004: 106)

In this sense, the occurrence of a Moorean belief implies that if a subject asks herself if she believes a Moorean proposition, she would find a contradiction. Such an approach takes it that the second-order belief is somehow inherent to the first-order belief. According to Kriegel, "Shoemaker can allow himself to construe the second-order belief as embedded in the first-order one precisely because it is only a dispositional belief" (Kriegel, 2004: 108).

The immediate criticism to this approach is that being a rational agent doesn't seem to require from someone to inquire about each of her beliefs. In a case where someone has a Moorean belief and does not ask herself about such a belief, nothing will preclude her from keeping believing it. In such a case, that person will not have the second-order belief that she believes a Moorean proposition. Therefore, she will not have actual contradictory beliefs (but only potential ones) (Kriegel, 2004: 107).

Shoemaker defends the principle that thinking limits speaking. According to this, it would not be possible for someone to utter a MS because it is not possible to believe a MS; at least, it is not possible to believe a MS without believing a contradiction. However, as Albritton (1995) accurately notices, if someone states a MS we would not say to her "no, you do not believe that because you just cannot believe that; it is impossible to believe such a thing". Instead, we would ask "Excuse me? What have you said?" This shows that Moorean sentences have another kind of problem.

In addition to such problems, Shoemaker's solution seems to be committed to a certain degree to the luminosity thesis, since he solves the paradox by relating first and second-order beliefs in order to indicate the contradiction present in a Moorean belief. That is, this kind of solution holds that the mere fact of having a belief that *p* allows me to have a belief about my belief that *p*. Cases of failure of

one's accessibility to his own mind, such as self-deception, indicate that the thesis of luminosity seems to involve serious problems concerning a proper conception of the mind and one's access to it.[14]

Despite these problems, Shoemaker's position carries important insights that we must not ignore. One of them is that the problem with a MS surpasses the act of enunciating it; it is not only a behavioral transgression. It is also and primarily a transgression about what is asserted and believed (Kriegel, 2004:113-114). The problem involves more aspects of the mental life of the subject.

All three solutions offer us important insights that we must take into account. However, I would like to offer a fourth one, inspired both by these insights and by the Davidsonian account of the phenomenon of communication. My proposal suggests that the different solutions are more useful when they are seen as complementary rather than as rival theses. Then, I will offer a new explication of why MS are absurd.

3. The irrationality of a MS

I have suggested that the first solution accommodates the intuition that a statement about the world shows us something about the person who states it. It highlights the kind of perspective that a person has, the place from where she is able to see the world, and the things she commits herself to. The second solution, I have maintained, is not just the inverse path –the idea that when someone says something about oneself she is saying something about the world– but also concerns the external aspect of the first-person perspective. From this perspective, questions about *p* and about my believing that *p* could be blended so that one question becomes just an alternative way of expressing the other. I take both ideas to be very reasonable

14 It is possible to argue against the luminosity thesis such as Williamson (2000) does, by offering a reductio ad absurdum proof.

and, for this reason, I will retain them within my proposal, through the following thesis:

> T1: the distinction between sentences of the type "*p*" and of the type "I believe that *p*" does not correspond to the difference between a statement that refers to the world and one that refers to the psychological life of someone, in the sense of making them independent from each other.

This thesis denies both that the general form of the type-sentences (whether "*p*" or "I believe that *p*") and their assertions in specific speech acts determine in advance that their contents are about the world or about the subject. T1 does not entail a stronger position that reduces one type of statement to another, as it occurs with the first two solutions discussed in the previous section. It is still possible to maintain that "*p*" and "I believe that *p*" are different statements. Nevertheless, T1 indicates that it does not make sense to identify the sort of reference each type of sentence has in an a priori manner; there is no reason to hold that sentences of the form "*p*" refer only to the world and sentences of the form "I believe that *p*" refer only to someone's psychological life.

In this sense, T1 rejects the presumption that a MS is a conjunction of two sentences about two different subject matters, logically independent from each other, though in conflict. Moreover, T1 doesn't intend to reduce one part of a MS to the other.

Regarding the third solution to the paradox, I will maintain a second thesis:

> T2: the absurdity of a Moorean sentence arises in speech as well as in thinking.

Actually, T2 does not answer completely to the third proposal because that proposal defends the primacy of the paradox at the level of beliefs over the level of assertions. Nevertheless, T2 retains the

intuition that the problem concerning a MS would still exist even if the MS were not uttered. T2 makes room for that intuition, but is guided by the idea that there is no sharp separation between the field of the mind and field of speech. So, regarding the third proposal, instead of maintaining the idea of primacy of the paradox at the level of beliefs, I will hold that the paradox also spreads to such a realm.

I will also develop a third thesis, but before that I will try to motivate the plausibility of T1 and T2. In order to do so, I will make use of Davidson's considerations about linguistic interactions, more specifically, of his ideas about radical interpretation and the theses involved in his understanding of it.

According to the Davidsonian view of radical interpretation, the attribution of mentality to someone and the attribution of meanings to her words occur simultaneously. That is, the attempt at understanding what someone says and the consideration of such a person as a subject with a mental life, do not occur in distinct moments. Let's think about the famous radical situation, where a translator seeks to elaborate a translating manual of a language used by a completely isolated society into English, for example. Underlying such a situation there is a sort of presupposition about the ability of the interpreter in recognizing intentional and linguistic behavior in the interpretee. Translating 'Gavagai' only makes sense as an activity if the interpreter understands that such a sound is emitted by a creature that copes with the world in a certain way that is very similar to our own, and that such a being wants to say something when she says 'Gavagai'.

The identification of intentionality, that is, the attribution of a mind to our interpretee, has to be part of the same task of understanding his words. If "Kurt utters the words 'Es regnet' and under the right conditions we know that he has said that it is raining" (Davidson, 1973: 125), we do this since by "having identified his utterance as intentional and linguistic, we are able to go on to interpret his words: we can say what his words, on that occasion,

meant" (Davidson, 1973: 125). The very activity of describing linguistic behaviors is already an activity of interpretation, in which we attribute not only meaning to someone's words but we also attribute mental states to him, such as the ones of believing and desiring.

It is necessary to notice that Davidson's view can be expanded, and in fact is expanded, from radical to ordinary communicative situations. According to Davidson (1973: 125), every case of understanding other's speech involves a radical interpretation that emerges for users of the same language, in terms of the question: How could we determine that the language is the same?

If this approach is plausible, it seems that both T1 and T2 get stronger. Someone that denies T1 seems to agree that it is possible to hear "*p*" without taking into account what does it mean for "*p*" to be uttered by someone. In the case of T2, its denial seems to involve the idea that thought and talk can be considered as two independent fields. Nevertheless, as I've suggested before, such contexts just increase the plausibility of T1 and T2. There is no direct consequence from Davidsonian positions to such theses. This is graphic when one realizes that in the case of a MS, we are dealing with particular sentences, while Davidson's account always deals with general theses about language and interpretation. If my perspective is correct, I did not demonstrate that "*p*" does not refer only to the world neither that "I believe that *p*" does not refer only to the speaker, but I believe that the Davidsonian account motivates the defense of such ideas. And more than that, it motivates us to be more holist concerning this matter. It seems to be possible that both "*p*" and "I believe that *p*" make explicit information about the world and information about the agent in the world, at the same time.

Defending the plausibility of T1 and of T2, nevertheless, is also insufficient to explain the kind of mistake a MS incurs in. In order to do so I present my third thesis:

> T3: A statement is absurd when it does not satisfy the principles of rationality[15].

T3 highlights and denies another assumption of the traditional debate about Moore's paradox: a statement is absurd only if there is a contradiction in it. It seems clear that a statement could be absurd due to several reasons. However, the paradox is traditionally understood only in terms of absence of a contradiction: "We know that a MS is absurd though there is no contradiction in the sentence itself"[16]. T3, especially in combination with T1, makes room for a solution that does without the reduction of one part of a MS to the other, which was needed to find a contradictory sentence behind a MS. The solution of the paradox via T3 is rather to explain why stating a MS cannot constitute a rational assertion.

Back to the same Davidsonian background about radical interpretation, the activity of interpretation is successful only if we attribute a certain degree of consistency and coherence, as well as a certain degree of correctness, to the interpretee's linguistic behavior. These ideas are part of the so called 'principle of charity', that Davidson develops not as being one possible strategy to interpret someone, but rather as the only alternative we have to interpret each other. In this sense, someone's statements, in order to be considered as meaningful and intentional, can only be interpreted if we take such statements against a background of coherence and consistency. I will defend that besides this, we need to assume the personal unity of our interpretee. I will call this set of assumptions we make when we interpret someone the 'presumption of rationality' for our interpretee.

Let's think about the imaginary case that opened this text, where my friend says "it is raining but I do not believe so". As I

15 The principles of rationality which T3 refers to are broader than the notion of coherence and logical consistency. I will deal with this issue later on.

16 Williams (2006), for example, defines the paradox in those terms.

designed the case, we were both in front of a window through which we could see clearly that it was raining. I could try to reinterpret her words in other terms, for instance, as being a metaphor or even as being an alternative form of saying something else such as "I cannot believe it is raining! It has ruined my plans to go to the beach"[17]. But if I do not take this path, I must consider my friend as having an irrational behavior. I know that she is wrong when she says she does not believe it is raining, but I am unable to convince her about her error because she first said "It is raining". Has she made a mistake or does she know that it is raining? It seems that in order to understand her, I would have to divide her into two persons, one that affirms the first part of the statement and the other that affirms the second one. However, I could not see her as a divided person without, at the same time, seeing her as irrational.

Davidson (1982) presents an interesting account of irrationality, and takes it to be a phenomenon that occurs "inside the house of reason" instead of outside the rational realm. According to his account, irrationality is not a question of someone's action being unreasonable under my eyes; it is instead about the beliefs and actions that result from a failure inside the mental realm of the same person. According to Davidson's definition (1982: 179), cases of irrationality are those where "there is a mental cause that is not a reason for what it causes". It is only in this sense that we could say that "a person is irrational if he is not open to reason".

17 In that case, her statement would not be a genuine case of MS. In fact, there are ordinary examples that have the same form of a MS without being one, because they are able to transmit an intelligible meaning. For instance, the expression "I do not believe that honest politicians exist, but they exist" can be perfectly used to convey one's perplexity about the existence of so many corrupt politicians. Another ordinary sentence could be "São Paulo is a very unpleasant city, but I do not believe so", which could be used to express the idea that although most people consider São Paulo unpleasant, for me it is not so. It is interesting to notice that such sentences are very similar to MS, but when we use them with an intelligible sense, they could not be understood as being a Moore's case.

I consider this approach to be very useful for dealing with cases of irrationality because, on the one hand, it is broader than other views which consider irrationality as being failures of predetermined rules of reasoning and, on the other, such an approach treats cases of irrationality without transforming them into another kind of phenomenon; that is, without eliminating the irrationality itself[18].

Davidson's definition of irrationality in terms of mental states which are caused by other states that are not their reasons, pushes him to consider that "if we are going to explain irrationality at all, it seems we must assume that the mind can be partitioned into quasi-independent structures" (Davidson, 1982: 181). Davidson argues that it is only this way that a mental state could cause another without being its reason. Nevertheless, he also affirms that such an idea does not require that we enumerate or name the different parts of the mind. Neither there is any need to establish a sharp division between those parts. We could keep this discussion at a more abstract level. Davidson points out that even "phrases like 'partition of the mind' and 'part of the mind' are misleading if they suggest that what belongs to one division of the mind cannot belong to another" (Davidson, 1982: 181).

So far, I have suggested that an authentic MS should be understood as a case of irrationality because it makes evident the internal division of the person who states it: in the given example, my friend is in the world partly considering that it is raining and partly considering that it is not.[19] This makes my friend inconsistent,

18 Several accounts about irrationality end up dissolving the very phenomenon of irrationality. Davidson (1982) highlights that the Plato principle, which is based in the doctrine of pure rationality, is an example of this. Such an account dissolves the irrationality by transforming it into a case of ignorance. At the opposite side, according to the Medea principle, an irrational act would be the resultant of forces beyond the very person. Under this view, in a case of akrasia, the person would act against her best judgment because some strange force exceeds her will. This implies that akratic acts will not be intentional at all.

19 Moran (2001: 69) suggests that cases like akrasia and self-deception, that are

but not her words themselves.

I have maintained that, under the conditions of radical interpretation, if we want to interpret someone, we should assume the rationality of our interpretee. And such an assumption involves the attribution of a background of truth and coherence to her words and, in addition, the supposition that there exists a unity in her mind. In this sense, the Davidsonian principle of charity is here expanded because mere coherence (of the belief system, as well as the coherence between what is believed and what is said), or even the presumption of truth for someone's words, do not completely account for the rationality involved by personal unity. The Davidsonian explanation about irrationality pushes us to conceive the mind as divided into different parts, but this does not mean that we have to consider such a division in the moment of the interpretation. On the contrary, we cannot do it. Also according to Davidson, there is no doubt that the principle of charity in interpretation is opposed to the partition of the mind (1982: 184). That is, irrationality is explained in terms of the possibility of the mind's partition, but given that we do not take our interpretee to be irrational in the first instance, we do not have

cases traditionally taken as failures of rationality, can be understood in terms of Moore's paradox, exactly because they make room for situations of division of how one understands oneself: the division between an attitude I have and another attitude I attribute to myself. Moran offers the following parallel: "I know that what I am feeling must be jealousy (I believe it is raining), but there is nothing to be jealous about" (but It is not raining). In this case we do not find a contradiction because the sentence represents a possible state of affairs, but we know that there is a tension in that.

Moran has a very interesting approach to the way we access ourselves. One of his theses is that we access ourselves through two manners: from a first-person perspective and from the third one. This does not mean that we have two different objects accessible from each of those modes of access. They allow us to correct our own beliefs through the comparison of these two perspectives. In a case of self-deception, the subject would be irrational if he insists in sustaining two conflicting points of view. The possibility of two modes of access to ourselves does not give us the right to sustain both conflicting points of view at the same time.

to consider her as divided.

The existing problem in a MS could be formulated in the following terms: in a radical interpretation situation we just cannot begin by attributing a MS to our interpretee. For a MS is an expression of an irrational behavior. As I've tried to defend, we have reasons to accept that interpreting is only possible if we assume that we are dealing with a rational creature. It is clear that we have irrational behaviors and it is indispensable to reserve a space in our theories to deal with such behaviors. However, the important thing is that those behaviors cannot be the rule. In the example of the rain, if I was dealing with a totally unknown language, I just could not begin my interpretation by attributing something like a MS to my friend's words.

We are forced to take others' words as having something to do with their position in the world and as being the manifestation of how a person in particular deals with it. Moreover, we take for granted that such a person is not divided in several parts that interact independently between each other with other people and with the world. It is because of this that T3 gains more plausibility. A MS is problematic not because there is some contradiction in stating it, but because our notion of rationality presupposes personal unity. We cannot attribute a MS under the principles of rationality because a MS hurts our logical intuitions, but more so because it hurts our self-conception as persons. And one of the essential aspects of this conception is that there is a unity in each of us; an idea that is totally different from the notion of total accessibility to our mental states and that is not a proof against the fact that we have contradictory behaviors or that we have tensions inside ourselves.

Theses T1, T2 and T3 together establish a context where we can understand the peculiarity of a MS with reference to the way we are committed to the idea that we interact linguistically as rational beings. In this case, acting linguistically does not involve that the field of speech and the field of thinking are detached. Neither it

can be assumed that someone can be neutral when she speaks about herself or about the world; in both situations several traits about how one thinks and about how one sees the world are revealed. If one says something that is only intelligible by considering her as being two persons, it is very likely that she is stating a MS.

Bibliographical References

Albritton, R. 1995, "Comments on 'Moore's Paradox and Self-knowledge'", *Philosophical Studies* 77 (2/3), pp. 229-39.

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

Bensusan, H. & Pinedo, M. 2007, "When my Own Beliefs are not First-personal Enough", *Theoria* 58, pp. 35-41.

Black, M. 1952, "Saying and Disbelieving", *Analysis* 13, pp. 25-33.

Davidson, D. 1973, "Radical Interpretation", en Davidson 1984, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press, pp. 125-39.

Davidson, D. 1982, "Paradoxes of Irrationality", en Davidson 2004, *Problems of Rationality*, Oxford, Clarendon Press, 169-88.

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

Heal, J. 1994, "Moore's Paradox: A Wittgensteinian Approach", *Mind* 103, pp. 5-24.

Hintikka, J. 1962, *Knowledge and Belief*, Ithaca and London, Cornell University Press.

Kriegel, U. 2004, "Moore's Paradox and the Structure of Conscious Belief", *Erkenntnis* 61, pp. 99-121.

Lee, B. D. 2001, "Moore's Paradox and Self-Ascribed Belief", *Erkenntnis* 55, pp. 359-70.

Linville, L. & Ring, M. 1991, "Moore's Paradox Revisited", *Synthese* 87, pp. 295-309.

Moore, G. E. 1942, "A Reply to My Critics" en P. Schilpp (ed.), *The Philosophy of G. E. Moore*, La Salle, Open Court, pp. 535-677.

Moore, G. E. 1944, "Russell's 'Theory of Descriptions'" en P. Schilpp (ed.), *The Philosophy of Bertrand Russell*, Evanston, Northwestern University, pp. 175-226.

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

Shoemaker, S. 1995, "Moore's Paradox and Self-Knowledge", en S. Shoemaker 1996, *The First-person Perspective and Other Essays,* Cambridge, Cambridge University Press, pp. 74-93.

Williams, J. N. 1979, "Moore's Paradox – One or Two?", *Analysis* 39, pp. 141-42.

Williams, J. N. 2006, "Wittgenstein, Moorean Absurdity and its Disappearance from Speech", *Synthese* 149, pp. 225-54.

Williamson, T. 2000, *Knowledge and its Limits*, Oxford, Oxford University Press.

Wittgenstein, L. 1953, *Philosophical Investigations* (translated by G. E. M. Anscombe), Oxford, Basil Blackwell, 1958.

6 Authority and Self-Knowledge

ABSTRACT
My aim in this paper is to develop a reading on first-person authority that dissolves the supposition that first-person authority is fully explained by the identification of the special traits of avowals. I argue that authority is rather a legitimate attribute of a person insofar as she is the only one able to have first and third-person perspectives on herself. It is because we are able to contrast both perspectives that we can legitimately be seen as an authority about what we think, even though we remain fallible. My proposal combines a non-epistemic approach to first-person perspective with an epistemic explanation of first-person authority. On the one hand, I defend a hybrid model of expressivism that incorporates transparency. On the other hand, I defend that the root of first-person authority is in fact epistemic.

KEYWORDS: first-person authority, self-knowledge, first-person perspective, expressivism, transparency.

RESUMEN
Mi objetivo en este trabajo es ofrecer una lectura sobre la autoridad de primera persona que disuelva la suposición de que ésta se explica totalmente por medio de los rasgos especiales de los *avowals*. Argumento que la autoridad es más bien un atributo legítimo de una persona en tanto que

ella es la única capaz de tener perspectiva de primera y de tercera persona sobre sí misma. Es porque somos capaces de contrastar ambas perspectivas por lo que se nos considera legítimamente como una autoridad sobre lo que pensamos, aunque seamos falibles. Mi propuesta combina un enfoque no epistémico de la perspectiva de primera persona con una explicación epistémica de la autoridad de primera persona. Por un lado, defiendo un modelo expresivista híbrido que incorpora transparencia. Por otro lado, defiendo que las raíces de la autoridad de primera persona son de hecho epistémicas.

PALABRAS CLAVE: autoridad de primera persona, auto-conocimiento, perspectiva de primera persona, expresivismo, transparencia.

Introduction

Self-ascriptions of mental states present special traits concerning its epistemic security. There is usually a presumption that the speaker is not mistaken when she avers her beliefs, hopes, desires, intentions or feelings[1]. When, for example, I tell a friend that "I'm excited about going to the mountains", unless she has reasons to doubt about my sincerity or unless she knows me well enough to remind me that I don't appreciate mountains that much, my claim about my present state of mind will be out of doubt. After all, who could tell better than me how I feel about my planned trip? Maybe my friend could, but the phenomenon of first-person authority is revealed by the fact that in most cases, what I say about my thoughts goes. That is why investigating the basis of such an authority has turned out to be a

1 This is approximately how Davidson (1984: 3) identifies first-person authority. I've added "feelings" to the group of mental states enumerated by Davidson, as representing the so called 'phenomenal avowals'. However, I will not endorse the division sustained, for example, by Crispin Wright (1998) between what he calls 'attitudinal' and 'phenomenal' avowals. I will focus instead on the division between avowals (in general) and non-avowals self-ascriptions.

promising starting point in order to understand the major question of self-knowledge.

It has been supposed that such an authority finds its legitimacy in the special mode we know our thoughts; a mode that is typically first-personal. We can know our thoughts by a similar manner to how others know them. Going to a psychoanalyst, talking to a friend or even reflecting on one's own behavior involve third-person strategies to access one's mind. However, there seems to be another way of knowing one's thoughts that is exclusive to the very person that entertains such thoughts; a mode which produces immediate and authoritative instances of self-knowledge, sometimes called 'avowals' (see Wright 1989, Bar-On 2004 and Moran 2001). Therefore, it has been supposed that among the several mental states a person ascribes to herself, only avowals should legitimately present first-person authority and at the same time, explain it. In such a view, the explanation of authority lies in the explanation of avowals.

My aim in this paper is to develop a reading on first-person authority that dissolves such a supposition; the supposition that first-person authority is fully explained by the identification of the special traits of avowals. I argue that authority is rather a legitimate attribute of a person insofar as she is the only one able to have first and third-person perspectives on herself. It is because we are able to contrast both perspectives that we can legitimately be seen as an authority about what we think, even though we remain fallible. Explaining first-person perspective is still a crucial question within the whole enterprise of understanding self-knowledge, but it doesn't explain by itself first-person authority. Our interlocutors have no clue to know whether our self-ascriptions are avowals or mere products of self-reflection. But the point is that attributing authority doesn't require such an ability once authority doesn't lie on avowals' characteristics. My utterance "I'm terribly tired" could be an immediate and spontaneous claim as well as the product of the recognition of my tired face being reflected in a mirror. However, it seems reasonable to

say that I don't lose my authority in the second case.

My strategy will be, first, to argue for an account of self-knowledge based on the necessity of both perspectives on oneself, the first and the third one. I begin by exposing a dilemma between the Cartesian and the Rylean approaches, which characterize self-knowledge exclusively in terms of one or another perspective. I will defend that both sides of the dilemma fail for not giving space to an accurate image of self-knowledge. I will offer then a summarized discussion on some different models of first-person perspective. I will examine three of them: the detectivist, the constitutivist and the expressivist views. I will favor a hybrid version of a variation of expressivism (Finkelstein's model) and of Moran's account, although my explanation of first-person authority might be consistent with some other models.

Having defended the necessity of both first and third-person perspectives on oneself, I will propose a puzzle based on the mentioned supposition; the supposition that first-person perspective is able to provide the explanation of first-person authority. I will try to show that this supposition leads us to undesirable conclusions. The dissolution of such a puzzle will give place to my view that authority should be an attribute of a person instead of her avowals.

1. Descartes versus Ryle. The necessity of both perspectives on oneself.

Descartes and Ryle represent two historical paradigms in the investigation of self-knowledge, generally seen as opposed ones[2]. One face of the abyss between them can be revealed in terms of the dilemma that emerges from considering first-person and third-person perspective as two exclusive clues to characterize

2 Davidson (1984, 1987), Moran (2001) and Bilgrami (2006) identify the problem of self-knowledge as highly subsidiary of the debate between Descartes and Ryle.

self-knowledge[3]. While in Descartes' approach self-knowledge is completely first-personal, in Ryle's approach the opposite goes, third-person perspective is all we need.

By 'self-knowledge acquired from third-person perspective' I mean self-ascriptions of mental states made on the basis of external evidence, inference, analysis, or self-interpretation[4]. One's linguistic and non-linguistic behavior are usually the primary basis for reaching one's mental condition under this procedure. By 'self-knowledge acquired from first-person perspective' I mean self-ascriptions called 'avowals', which enjoy a sort of immediacy. Given the variety of explanations of first-person perspective, this very general definition is neutral with respect to the conception of avowals as epistemic or non-epistemic products.

Cartesianism[5] takes self-knowledge to be completely first-personal. Under this approach[6], first-person perspective to oneself is a cognitive process, which is inner, conscious, and has special epistemic qualities. The supposed inner realm is known by the subject through a kind of "inner sense", which is a direct, complete and infallible mechanism. His entire mind is illuminated to him.

3 Bilgrami (2006: 9) identifies this dilemma in the following terms: "So something like a dilemma has emerged: if one accepts the Cartesian assumption of subjectivity regarding the nature of mental states (and the meaning of mental terms), one has an insuperable problem of other minds, and if one rejects that assumption and adopts a more third person perspective on mental states, that leads –at first sight, anyway– to counterintuitive consequences for our knowledge of our own minds". I will argue rather that neither side of the dilemma is able to successfully explain the phenomenon of self-knowledge insofar as both perspectives are needed for one to be a cognitive agent with self-knowledge.

4 I'm using Bar-On's (2004: 226) characterization of third-person perspective.

5 It seems reasonable to investigate the question Moran (2001:12, footnote 9) indicates in a footnote: "Was Descartes himself a Cartesian […]?" However, following my election in the preceding chapters and given an apparent agreement in talking about Descartes as committed to the doctrines of infallibility and self-intimation. I will suppose an affirmative answer to such an inquiry.

6 Cartesianism is part of a more general model called by Finkelstein (2008) 'old detectivism'. I will return to it in the next section.

The analogy between the alleged inner sense and the external senses ends up with the requirement of infallibility. There is no parallel of such characteristic which can be possibly found in perception. Self-knowledge is completely first-personal because there is no better perspective to take over oneself than one's own.

Once everything there is to know about one's mind is accessible to the subject, without space to misunderstandings, authority is automatically explained. I am correctly attributed with first-person authority because the cognitive process involved in first-person perspective guarantees the best epistemic position to the subject. My unique perspective on me is such that enables me to know all my thoughts directly and infallibly. Assuming a third-person perspective on me would be useless or even prejudicial. The authority attributed by others to my statements about my mind reflects and is explained by my epistemic privilege over myself.

This model explains both the special character of first-person perspective and the authority with which one thinks and speaks about one's mind. However, the very characteristics that explain authority lead self-knowledge into a collapse. It seems clear that not only we have some mental states hidden from us, but also that we are sometimes mistaken in our judgment about them. This is noticeable when, for example, a close friend helps us to know that what we say we believe –e.g. "I find nationalism repugnant"– conflicts with our behavior –e.g. I almost die when my country team is playing on the World Cup and I voted for the criminalization of illegal immigration. Such ordinary facts are clearly underestimated by Cartesianism. Completeness and infallibility are definitely not adequate traits of self-knowledge.

Self-ignorance and self-mistake seem to defeat Cartesianism[7].

7 Another important criticism towards Cartesianism concerns the conception of the mind as an inner space composed by inner objects and the conception of self-knowledge as an observation of this inner realm. I will comment this criticism, especially under Wittgenstein's arguments, in the following section. In a sense,

Besides not taking into account those common cases in its original formulation, if Cartesianism were to incorporate ignorance and mistakes, it would need to abandon the exclusiveness of first-person perspective. For it seems that third-person perspective has a crucial role in the revision of one's own thoughts. Suppose a friend tries to convince me that I believe there is someone who pursues me while I deny it. I'm able to change my mind only if I can follow her by a third-person mode of reasoning. She can bring my attention to my behavior that does indicate that I believe there is someone pursuing me: I'm constantly looking back and showing an expression of being afraid; I'm walking faster than usual all day long. I'm able to change my mind and convince myself that in fact I believe there is someone pursuing me, because I'm able to occupy a position of another person over me. I'm able to understand my friend's reasons and reflect on my behavior somehow by distancing myself from my usual perspective. And in this case, only by doing this, I reach self-knowledge.

Let's return to the case where someone states "I find nationalism repugnant" while her ordinary behavior indicates the contrary. And let's suppose she knows well enough what nationalism means. It seems that she will not change her avowal if she doesn't listen to what her friend says. Suppose her friend insists "No, you don't find it repugnant, in fact you are quite nationalist. Look, for example, how you support your national team and how you voted the question of immigration". No matter how many times she asks herself what she believes, she will not get an accurate response unless she analyzes herself. In other words, no matter how much she insists on her first-person perspective, there is no way to change her mind unless she brings some pieces of information into it. It is not just a matter of refocusing the attention. Sometimes it is, but not always.

Ryle's proposal is also a profound reaction to that idea, although his view remains as problematic as what he criticizes.

Suppose I say to my partner "I'm terribly irritated with this crowd" and he corrects me "no, you are not, you are just hungry". I can immediately realize that I'm hungry and, still from a first-person perspective, say "it is true, I'm terribly hungry; that is probably why I feel irritated". However, it is difficult to imagine that this could happen in the nationalism example and other several cases. Even if we are to consider third-person perspective as a moment of passage towards a *personalization* of a self-ascription, this perspective seems to have an indispensable role in such cases[8].

Therefore, first-person perspective alone doesn't seem to be sufficient to characterize a real subject that has self-knowledge and who can revise one's thoughts. But, at this moment, one could wonder whether it would be possible to defend the other extreme position: to consider self-knowledge as entirely acquired by the same kind of reasoning others take over me. That is, beyond necessary, could we say that third-person perspective is sufficient for self-knowledge?

A positive answer is given by Ryle. Others' knowledge about my mind and my self-knowledge are different in degree, but not in kind (Ryle, 1949: 179). There are no differences in terms of methods which I and others employ to know my thoughts. There is no asymmetry between first and third-person perspective insofar as the first one is a sort of illusion. The usual epistemic familiarity I have with my thoughts lies on the fact that I'm constantly receiving

8 Children seem to provide good indications of the importance of third-person perspective in calibrating first-person perspective. When, for example, children become irritated because they are tired, it is not unusual to see that some of them insist on asking for anything else but going to bed. They claim to be angry because they want to see another cartoon, to play more games, etc. In a sense, they are not yet completely able to avow how they really feel, that is, tired. It seems that the incorporation of the third-person perspective (through accepting parent's attribution of desires and thoughts, for example) is what partly allows children gradually to distinguish among their states of mind. However, this needs further argumentation.

data from my behavior. I cannot help having me among the portions of the world I pay attention to. And, first-person authority can just respond, at the most, to such a contingent fact. I am in a very good position to know what others are often in a very poor position to know (Ryle, 1949: 179). Because of this, it is possible to maintain the idea of epistemic superiority of the speaker over the listener. However, this should not indicate that the former has any privileged access[9] to one's own mind, such as the one envisaged by Cartesianism.

The strategy Ryle develops to sustain his view involves the dismantlement of the Cartesian theory, identified as the 'official doctrine'. According to Ryle, such a theory is committed to a profound mistake, a "category-mistake"; it involves the representation of "the facts of mental life as if they belonged to one logical type or category (or range of types or categories), when they actually belong to another" (Ryle, 1949: 15). The specific mistake involved here is to consider mind-vocabulary and body-vocabulary as terms of the same logical category. That makes room for what Ryle frequently refers to as the 'dogma of the Ghost in the Machine'. Ryle argues that we should not understand the mind as something beyond the person himself[10].

9 The notion of privileged access is usefully examined by Alston (1971). According to him, this notion is committed to the idea that self-knowledge is fundamentally different and also superior to others' knowledge about one's mind. However, there are several different explanations of such an epistemic superiority. In brief, Alston (1971: 239) identifies the following principles as the possible explanations: i. Infallibility; ii. Omniscience; iii. Indubitability; iv. Incorrigibility; v. Truth-sufficiency (a weaker analogue of omniscience); vi. Self-warrant (a weaker analogue of infallibility). Alston (1971: 240) argues that "attacks on privileged access invariably fail to take account of the full range of possibilities" and he defends notion (vi) as the most defensible principle. In the above passage, my use of the term 'privileged access' refers to the notions (i) and (ii), which I attribute to Descartes. However, this is the only passage where I use such a term, since it doesn't give space to non-epistemic views, which are successfully treated under the term 'first-person perspective'.

10 It is important to notice that Ryle makes it hard to side him either with

Once such a mistake is dissolved, we are able to dissolve also what Ryle calls the 'phosphorescence theory of consciousness'[11], the theory involved in the Cartesian account of self-knowledge. He will conclude that there is nothing mysterious or occult in our self-ascription of mental states. Those self-ascriptions are the same in kind as the mental states others attribute to me, as is the method applied by both of us to make such ascriptions. We know about other people and about ourselves through paying notice to behavior, linguistic or not. Ryle indeed recognizes that self-ascriptions are generally taken as primary source of information about a given person (Ryle, 1949: 181).

Based on Ryle's approach, one could possibly explain first-person authority though in a very different line from Cartesianism[12];

---

reductionism or with eliminativism: "If my argument is successful, there will follow some interesting consequences. First, the hallowed contrast between Mind and Matter will be dissipated, but dissipated not by either of the equally hallowed absorptions of Mind by Matter or of Matter by Mind, but in quite a different way. For the seeming contrast of the two will be shown to be as illegitimate as would be the contrast of 'she came home in a flood of tears' and 'she came home in a sedan-chair'. The belief that there is a polar opposition between Mind and Matter is the belief that they are terms of the same logical type" (Ryle, 1949: 23). In that sense, the Rylean rejection of the myth of the ghost in the machine involves not only the dissolution of the ghost, but also of the machine.

11 Ryle defines the phosphorescence theory of consciousness as committed to two main theses: (i) the contents of the mind are self-luminous, and (ii) the mind can "see" or "look" at its own operations in the "light" given off by themselves (Ryle, 1949: 159). These are two characteristics of the Cartesian model treated in the preceding pages.

12 Ryle appeals to two main conditions in order to explain the best epistemic position the subject has. First, there is a sort of supposition that presence guarantees knowledge. I know myself better because I cannot help being with me. This is graphic when Ryle says that "the turns taken by a man's conversation do not startle or perplex his wife as much as they had surprised and puzzled his fiancée" (Ryle, 1949: 179). He is probably supposing that a wife, having stayed longer with her husband, compared to the time they were only dating, should know him better. I would rather go slowly. It is not strange to recall some cases of couples who, living together only by routine or by any other practical reason, have lost completely interest in each other. In such cases, no matter how much

one that dispenses with appealing to the subject's special perspective. There is a compelling intuition behind Ryle's project, which is to dissolve both the useless idea of a "metaphysical Iron Curtain" that compels us to be absolute strangers to one another and the naive picture of a "metaphysical looking-glass" that compels us to be completely disclosed and explained to ourselves (Ryle, 1949: 181). However, in the middle of this path, Ryle has also supposed an unnecessary commitment: that first-person perspective is to be understood necessarily under the scope of Cartesianism. Let's consider the following case Ryle proposes:

> I come to appreciate the skill and tactics of a chess-player by watching him and others playing chess, and I learn that a certain pupil of mine is lazy, ambitious and witty by following his work, noticing his excuses, listening to his conversation and comparing his performances with those of others. Nor does it make any important difference if I happen myself to be that pupil. I can indeed then listen to more of his conversations, as I am the addressee of his unspoken soliloquies; I notice more of his excuses, as I am never absent, when they are made. On the other hand, my comparison of his performances with those of others is more difficult, since the examiner is himself taking the examination, which makes neutrality hard to preserve and precludes the demeanour of the candidate, when under interrogation, from being in good view. (Ryle, 1949: 169)

---

time they spend together, living in the same house, eating at the same table or even sharing the same bed, they will not know each other better than a close friend would. They have clearly lost their real possibility of knowing each other's thoughts. They just don't care. Maybe in the case of oneself, it is difficult to assume such a posture of not caring for oneself. However, Ryle still seems to require an additional argument. The second condition, also very questionable, is maintained by Ryle with reference to the fact that we normally take for granted other self-ascriptions even knowing how often people keep things back, are "insincere" or tendentious. Ryle argues that the normal, or "natural situation", is one where the speaker speaks one's mind. To refrain from doing this would be the "sophisticated" course of action (Ryle, 1949: 181).

However, the difference between being and not being that pupil is precisely the one between having and not having a first-person perspective. Suppose that the pupil, Emma, is training with her partner a determined sequence of strategies. Between one move and another, she states "It is useless to follow this strategy; we can predict exactly the next move each of us is going to perform. This is not fun". At the same time, Emma directs her attention to her behavior in order to know what she is thinking. She notices that she keeps following the predicted moves; she appears to be excited about playing and anxious about beating her friend this time. Emma concludes, though: "I believe it is useful to follow this strategy; I believe we cannot predict exactly the next move each of us is going to perform. I believe this is great fun". To sum up, Emma is ready to state one instance of a Moorean sentence, such as "this is not fun, but I believe it is fun"[13].

The intuitive reaction, however, would be to say that her first verdict about the game should be reflected on what she herself thinks. "She just cannot judge that playing chess with a certain strategy in mind is not fun while she believes it is fun[14]!" One would like to say. "She must be committed to her own judgments!" From the third-person perspective, we can perfectly perceive another person's beliefs in discordance with the way the world is, e.g., "Emma believes chess is a terribly boring game, but chess is in fact very fun". However, it seems that from the first-person perspective, if the beliefs in

13 This is a version of Shoemaker's argument (1994) against observational models of self-knowledge, which accuses them of being consistent with a person failing to have first-person perspective –the self-blind person. He argues against the intelligibility of such a person. I will return to this argument in the next section.

14 The word 'fun' may suggest a subjectivist interpretation of the term; one that supports a reinterpretation of the sentence such as "people think this game is not fun, but I don't agree with that". I don't intend to use the example in this sense. The given example deals with a real judgment about the game given by Emma: "it is not fun".

discordance are mine, I cannot maintain them[15]. I'm judging them as mistaken!

This is the underlying intuition behind the notion of transparency[16], which seems to indicate an important mark of first-person perspective. When we assume a first-person perspective, questions about the world are transparent with respect to questions about what the subject believes. That is, from a first-person perspective, "the first-person question 'Do I believe P?' is 'transparent' to, answered in the same way as, the outward-directed question as to the truth of P itself" (Moran, 2001: 66). Although they have different truth conditions, the subject answers to both in similar ways. For, in the normal case, when I ask myself about my belief whether *p*, I direct my attention outward, not inward. This

15 As far as I can see, it is possible for someone to state a real Moorean sentence. However, such a behavior is as irrational as any piece of irrational behavior we find ourselves performing, even if they are very rare. In Borgoni (2008) –Chapter 5 of this dissertation– I argue that if the interlocutor doesn't reinterpret the Moorean sentence just said, her only alternative option of getting some sense from the speaker's words is by dividing the asserter into two persons, one who affirms the first part of the sentence and another who affirms the second one. In that case, however, one could not see someone as a divided person without, at the same time, seeing her as irrational.

16 One terminological remark is important here. Transparency and luminosity, as I'm using them, are two very different notions. I'm using the term 'luminosity', which appears in the literature also under the name of 'self-intimacy', as referring to the thesis that the entire mind of a subject is accessible to him. In other words, any first-order thought the subject entertains is potentially knowable to him. Luminosity is a notion that has appeared along the exposition of the Cartesian account. I'm using the term 'transparency' to refer to a very different relation also between first and second-order judgments. It refers to the kind of reasoning one subject entertains when she assumes first-person perspective: while deciding about her second-order thought, the first-order thought comes to light. Because of this, those notions are not the same, nor represent opposite relations. It seems indeed that both are compatible, as seems to be the case present in Shoemaker's account, who defends a weaker (non-Cartesian) version of self-intimacy (Bar-On, 2004: 407 footnote 3). For more on luminosity see Shoemaker (1996, 1994) and for a proof against luminosity, see Williamson (2000). For more on transparency, see Moran (2001), Evans (1982) and Wright (1998).

insight –suggested, among others, by Evans (1982)– completely dissolves the Cartesian picture of self-knowledge as a privilege view towards inner states and happenings only accessible to the person herself (Evans, 1982: 225). Moreover, it reveals one particularity of first-person perspective: the commitment that one has to one's own judgments. This is so because the intelligibility of my answering questions about what I believe by focusing my attention on the very object of my belief is only possible if my own belief is, in some sense, "up to me" (Moran, 2001: 66).

Ryle's account, however, lacks the tools for explaining such a particularity; for explaining transparency, which seems to be an important mark of avowals, in contrast with non-avowals self-ascriptions. In addition, insofar as there is no such a thing as first-person perspective, it seems that Moorean sentences are exactly of the same kind as ordinary sentences such as "it is raining but Emma believes it is not". However, uttering a Moorean sentence is not rational at all. And transparency seems to give the supposed unity one expects of one's judgments, that somehow prevents one from affirming sentences such as "it is raining but I believe it is not".

Besides this class of arguments that refers to a distinctive mark of first-person perspective, it is possible to defend the necessity of such a perspective through another class of arguments, which refers to the requirements for rationality. Shoemaker (1994: 285-86) for example, finds first-person essential for revising one's thoughts. In the preceding lines, I've defended that revising one's own thoughts demands being able to take a third-person perspective on oneself. However, it seems that revision also demands first-person perspective insofar as one needs to be able to identify which thought is under revision in the first instance. Burge (1998b), following a slightly different argument, argues against a position which maintains that we can fully understand reason or thought without making use of the first-person concept or without appealing to a first-person

perspective (Burge, 1998b: 249)[17].

According to Burge, reasoning involves not only abstract evaluation of attitudes or of relations between thoughts, but also practical implementation of them in actual reasoning. He maintains that someone can only understand reason if she is capable of making use of reasons and, furthermore, if she actually uses reasons to support or to modify her own mental states in her thinking practices. In reasoning, one is immediately moved by reasons (Burge, 1998b: 250). He insists that a thinker, when reasoning, cannot just be a passive spectator given that the very idea of having reason demands that reasoning possesses the capacity to be motivated by the power of reasons (Burge, 1998b: 250-51).

Burge argues that a full understanding of reason involves being capable of marking the distinction between cases where rational evaluation motivates immediate implementation in one's attitudes and cases where it doesn't. However, even for those thinkers who lack the conceptualization of such a mark, the use of first-person concept in self-attributions sets apart those acts and attitudes that rationally demand to be rationally evaluated in order to be changed or maintained. Such a concept sets the "the locus of responsibility" for epistemic and practical agency (Burge, 1998b: 253). In a sense, this is an alternative strategy of arguing for a distinctive aspect of first-person perspective[18], namely, that reasoning from that

17 Burge finds in G. C. Lichtenberg's remarks in *Schriften und Briefe*, ii (Carl Hanser Verlag, 1971, 412 §76) the starting point of his discussion for exhibiting the sort of position he is arguing against.

18 I'm supposing that we can transpose Burge's argument in terms of the first-person concept 'I' to the discussion on first-person perspective. This possibility seems to be allowed. Burge, for example, differentiates third-person attributions from first-person as following: "third-person attributions do not mark the immediate rational relevance of rational evaluation to implementation of the evaluation. Even when a third-person attribution is to oneself, the relevance is not rationally immediate. For one could fail to know that the third-person attribution applied to oneself. I could fail to know that I am Burge. And although I do know, the rational relevance of reasons to their affecting my attitudes is not

perspective involves "immediate implementation of the evaluation on the evaluated attitudes" (Burge, 1998b: 258).

2. Models of self-knowledge: varieties of first-person perspective.

So far, I have defended the necessity both of third-person and of first-person perspectives in order to understand self-knowledge. This point was my basic argument against Cartesian and Rylean approaches. However, while the definition of third-person perspective doesn't seem to involve deep disagreements, the characterization of first-person does. In what follows, I will discuss three models of first-person perspective that differ in quite sensible ways from Cartesianism.

Following the terminology offered by Finkelstein, Cartesianism could be located within what he calls the 'old detectivist model' of self-knowledge, which can also be found in the literature under the name of 'observational' or 'perceptual model'. As I have pointed out in the last section, such a model identifies first-person perspective as a cognitive process loosely analogue to external senses, yet it maintains very particular epistemic qualities: it is a complete and infallible mechanism. I argued that those traits are very problematic: we do have some states of mind that we ignore we have and some others that we are mistaken about. In addition, accounting for those situations seems to require the inclusion of third-person perspective; a necessary perspective to understand somebody else's statement about my thoughts and to correct my own judgments about them.

Nevertheless, one could try to save the detectivist spirit of this model by rejecting the requirement of infallibility (and possibly, of completeness). This maneuver leads us to what Finkelstein calls

conceptually immediate. It must pass through the assumption that I am Burge" (Burge, 1998: 255).

the 'new detectivist model of self-knowledge'[19]. This new version of the detectivist model preserves the idea that taking a first-person perspective over oneself is to engage in a cognitive process of discovery of inner items. However, inner sense, in this new version, is taken as a much stronger analogy to external senses. In this model, mental items are independent from the subject's consciousness of them. Moreover, the faculty of inner sense could fail by several reasons, either by malfunction or by misperception.

Armstrong (1968: 110) –one of the representatives of this model[20]– insists that in introspection as much as in perception, where we distinguish the perceiving and the thing perceived, it is necessary to set the introspection apart from the thing introspected. Although they are both mental states, they are not the same: "a mental state cannot be aware of itself, any more than a man can eat himself up" (Armstrong, 1968: 110). He adds that, once more like in perception, introspection can be erroneous and even incomplete. In the same sense that there are many features of our environment that we fail to perceive in any given perception, in our awareness of our own mind we also fail to be aware of many mental states and of many of their features (Armstrong, 1968: 111). Yet according to this model, even without infallibility, one, as compared to the others, still maintains a better epistemic position over one's own mind insofar as those items are internally located.

19 Bar-On refers to this model as the 'contemporary materialist versions of introspectionism' or simply as the 'materialist introspectionist'. She summarizes such a model as following: "On one story, it is speculated that the human brain is equipped with a special mechanism – a scanner – designed to deliver reliable higher-order judgments about our first-order mental states. My distinctive ability to tell what I am thinking right now, for instance, is due to my brain's ability to scan its own present operations so as to yield highly reliable, non-evidential judgments, which are then articulated (in speech or in thought) through self-ascriptions of mental states" (Bar-On, 2004: 96).

20 Finkelstein (2008) remarks that there is a great variety of positions under the same label of 'new detectivism'. However it seems that the traits here enumerated can be faithful for most of them.

The new detectivism, though, corrects some of the old version's serious mistakes by strengthening the analogy between inner sense – the cognitive process by which one allegedly knows one's mind– and external senses. This makes us able to introduce error and ignorance into the model. However, this very remedy makes room for another sort of criticism. If inner sense is similar to external senses, it incorporates the possibility that the mechanism collapses. Therefore, it is possible to conceive a person who completely fails to have such a sense: a self-blind person; peculiarly, the opposite risk to the one presented by the old version of detectivism.

The problem that arises with this model is the same that arose in Ryle's account. Once new detectivism is compatible with a self-blind person, the model as a whole is not able to maintain first-person perspective as a necessity. It is important to notice that the notion of self-blindness supposed by the argument doesn't entail any cognitive deficiency. Such as in ordinary blindness, the self-blind person is conceived as being in principle equal in intelligence, rationality and conceptual capacity to someone who is not self-blind (Shoemaker, 1994: 281). Such person is in fact a Rylean creature, who only has third-person perspective to herself. The problem is that such creature's thoughts fail to exhibit transparency, a distinctive trait of avowals, which has consequences to one's rationality. Transparency, for example, seems to prevent one from affirming a Moorean sentence. Without transparency, nevertheless, it seems there is no difference between affirming "It is raining but Emma believes it is not" and "It is raining but I believe it is not". Maybe, one could learn to avoid such sentences but she will still lack a very important characteristic related to transparency: the characteristic of taking one's thoughts as being one's own. This was the first argument presented for the necessity of first-person perspective, which is approximately how Shoemaker argues for the unintelligibility of a self-blind person. In addition, we have seen a second argument against the possibility of self-blindness: we need to have a range of unmediated knowledge of

our minds in order to be able to realize that some thoughts need to be revised and in order to be able to identify such thoughts.

However, there is still a third argument against the detectivist model. It applies both to the old and to the new versions of it. The criticism, emblematically developed by Wittgenstein, intends to attack the picture of self-knowledge as a cognitive activity of detection of inner objects. This sort of criticism has given birth to some reactions to the detectivist ideal of self-knowledge, such as the one traced by the constitutivist model.

The constitutivist model of self-knowledge, in absolute contrast with the detectivist model, doesn't take self-knowledge acquired from a first-person perspective as consisting of a cognitive product. There is no cognitive mechanism involved at all. When one thinks or asserts states about oneself, there is rather a moment of constitution of the very person's mental states. Also in contrast with the new detectivist model, there is no independency between the known objects and the subject's knowledge about them, since the subject participates in their very appearing. Avowals are, in a sense, like "performative acts which bring into existence the relevant states of affairs, acts of forming an intention, deciding what to want, believe, hope for, etc" (Bar-On, 2004: 141).

Crispin Wright, one of the proponents of this model, finds in Wittgenstein the necessary reasons both to deconstruct Cartesianism (and variations of the detectivist model) and to defend a new approach to avowals. According to him, the basic philosophical problem of self-knowledge is precisely to explain the phenomenon of avowals (Wright, 1998: 22). Wright divides avowals into two classes, 'phenomenal avowals' and 'attitudinal avowals'. The former involves self-ascriptions of sensations or feelings such as "I have a headache" and the latter involves states partially individuated by propositional content, such as "I believe the temperature is around 40ºC". According to Wright, both classes of avowals are groundless, exhibit transparency and are authoritative. However, while phenomenal

avowals are strongly authoritative, the attitudinal ones are only weakly authoritative. The latter are more open to doubts than the former. This division is important for Wright because he understands Wittgenstein arguments as a "two-pronged attack on the Cartesian picture". According to him, the conception of phenomenal avowals as inner observational reports is challenged by the private language argument, while the parallel conception of attitudinal avowals is subject to the criticism found in the "not a mental process" passages that can be found in many places through the *Investigations* (Wright, 1998: 25).

Wright's reading of the anti-private language argument and of the remarks on following a rule also gives the basis for his positive account[21]. The alleged paradox of PI §201, in which every course of action seems to possibly accord or disaccord with the rule (and so, empty the notions of agreement or conflict) is solved by Wright as an alternative to Kripke's solution (Finkelstein, 2008: 29). The Kripkean skeptical paradox is solved by Kripke through a skeptical answer: although there is no matter of fact where to base the truth values of a sentence such as "Jones means *plus* by plus", there is still a sort of correctness to be established within a community. The alternative solution Wright defends is one that Finkelstein calls 'stipulativism' of meaning, a position along the following lines: "the fact about my past usage of 'plus' that fixes it that I am now acting in accord with what I then meant by 'plus' is just that I meant plus by

21 See chapter 2 "Externismo sin Experimentos Mentales" for an extended discussion on Wittgenstein's arguments. There, I defend that both private language argument and rule-following considerations provide good bases for defending externalism. The fact that Wright, among others, identifies those same arguments as sustaining a non-Cartesian picture of self-knowledge seems to be indicative of my argument defended in Borgoni (2009a) –Chapter 4 of this dissertation– where I argue that externalism is compatible with self-knowledge except with the detectivist model. Since I take Descartes as representing both internalism and old detectivism, once Wittgenstein's argues against Cartesianism, he seems to provide a dual basis for criticism.

'plus' (Wright, 2001: 177 *apud* Finkelstein, 2008: 35). In summary, Wright interprets Wittgenstein as putting forward the idea that not only the content of rules, but also that of intentional states is "a matter for us to decide" [22] (Finkelstein, 2008: 37).

This specific reading of meaning will be reflected on the proposal of constitutivism, a view that Wright himself identifies as the 'default view' on self-knowledge. Such a view characterizes avowals as a moment of decision (or stipulation) instead of an act of report of inner states. When one self-attributes mental states, in thought or in speech, those very states are constituted at that moment. There is nothing inner being discovered. The inner is under constitution every time the subject engages in first-person perspective. And first-person authority takes such changes into account:

> [T]he authority standardly granted to a subject's own beliefs, or expressed avowals, about his intentional states is a *constitutive principle*: something that is not a by-product of the nature of those states, and an associated epistemologically privileged relation in which the subject stands to them, but enters primitively into the conditions of identification of what a subject believes, hopes, an intends. (Wright, 1989: 154)

Constitutivism has received some criticisms concerning its accuracy in characterizing avowals as well as concerning its accuracy in interpreting Wittgenstein. Bar-On (2004), for example, bases her criticism only on the first point. According to her, constitutivism is unable to explain the epistemic security of avowals in contrast with other non-avowals self-ascriptions. She reads Wright as indicating that such epistemic security is built by definition into the very truth-conditions of first-person mental ascriptions. However, there is nothing in such a conceptual constraint, inasmuch as it is applied to the mentalist discourse as a whole, that can help us to differentiate,

22 It is important to notice that the sentence "for us to decide" doesn't refer here to communitarism, but instead to estipulativism, as points out Finkelstein (2008).

not only between kinds of self-ascription, but even between self-ascription and ascription to others (Bar-On, 2004: 348).

Finkelstein (2008), apart from criticizing constitutivism as an appropriate model of avowals, also criticizes the interpretation of Wittgenstein that is on the basis of such a model. According to Finkelstein, Wright finds "stipulativism" in Wittgenstein where he should find a more radical view on meaning and understanding. And the second criticism made by Finkelstein accuses constitutivism of misrepresenting the subject's responsibility. To put it bluntly, if the relationship between having a headache and stating that one has it were so intimate, it would not make sense to feel sympathy for someone claiming to be suffering from a headache. In fact, it would make sense to blame the sufferer for her headache and tell her to stop going around saying that her head is aching. This is highly counterintuitive: the avowal of a headache is not what makes it awful[23] (Finkelstein, 2008: 47, 52).

One alternative model that emerges from both of those criticisms is expressivism. Some passages of Wittgenstein's writings seem to be

23 Finkelstein's criticism of McDowell's model can be put in similar lines: what is awful about headaches is not the passive actualization of phenomenal concepts, but the headaches themselves. Finkelstein (2008) identifies McDowell's view on first-person perspective as a 'middle path' between detectivism and constitutivism. McDowell characterizes 'inner sense' following his general model on perception, where experience involves conceptual capacities although remaining passive: perceiving my own mental states involves both receptivity and understanding. Experiences, "both inner and outer, are constituted by the actualization of conceptual capacities" (Finkelstein, 2008: 64). Differently from detectivism, McDowell maintains that the objects of inner sense are not independent from my awareness about them; that is, there is a moment of constitution in perceiving the inner. And, differently from constitutivism, there is something more fundamental than avowals themselves. However, it seems that the McDowellian formula doesn't avoid Finkelstein general criticism to constitutivism (which is not Finkelstein's criticism to McDowell): if pain, for example, is the actualization of conceptual capacities, and if we want to avoid painful states, it seems that the only thing to do is to avoid applying painful concepts, or primarily avoiding to learn new painful concepts.

indicative of this alternative option to Wright's interpretation, such as the following excerpts from *Philosophical Investigations*:

> 244. […] "So you are saying that the word 'pain' really means crying?"--On the contrary: the verbal expression of pain replaces crying and does not describe it. (PI §244)

> 585. When someone says "I hope he'll come"--is this a *report* about his state of mind, or a *manifestation* of his hope?--I can, for example, say it to myself. And surely I am not giving myself a report. It may be a sigh; but it need not. If I tell someone "I can't keep my mind on my work today; I keep on thinking of his coming"--*this* will be called a description of my state of mind. (PI §585)

Or the following passage from *Zettel*:

> 472. Plan for the treatment of psychological concepts.
>
> Psychological verbs characterized by the fact that the third person of the present is to be verified by observation, the first person not.
>
> Sentences in the third person present: information. In the first person present: expression. (Not quite right.)
>
> The first person of the present akin to an expression. (Z §472)

What Bar-On identifies as the 'simple expressivist account' finds its routes in those lines of thinking. This model understands avowals as being purely expressions, in contrast to the notions of report and of constitution. Taking first-person perspective is neither a moment of detection of those states, nor a moment of constitution. It is instead like "natural expressions". For those who defend such an account, there is little in common between my sincerely saying that I am in pain and someone else's claiming that I am in pain, given that my use of the words could perfectly be replaced by other kind of vocalization or even by facial expressions (Bar-On, 2004: 228).

Identifying avowals as purely expression has, however, a serious problem concerning semantic continuity, that is, the continuity between my avowal "I am in pain" and others' ascription "she is in pain". For it seems that one could perfectly use my avowals to describe

my state and to make inferences. Both are true or false inasmuch as they both identify the same individual and ascribe to her the same condition at the same time (Bar-On, 2004: 9)[24]. Once the simple expressivist account considers expressions as opposed to descriptive reports, avowals lack truth values. As I said, they are just as groans, they only suggest a person's present state of mind. However, it seems reasonable to say that if someone hears some avowal of mine, she will successfully describe my mental state by using my own words and she will be able to make inferences about myself. In this account, avowals could not serve as legitimate premises in logical inferences as they in fact do. Neither could they be something that the subject knows. However, there are alternative expressivist explanations that are not committed to the opposition between expression and report, such as the models defended by Bar-On or by Finkelstein.

Finkelstein, for example, bases his account on another interpretation of Wittgenstein's writings, defending that Wittgenstein does not force us to see avowals as expressions and, hence, not as assertions (Finkelstein, 2008: 99). He defends that the whole lesson to be learnt from the discussion on following a rule is that there is no gap between words and their meanings (or between intentions and actions, etc). Finkelstein points out that the paradox of PI §201 has its roots in a misunderstanding suggested by PI §431: "There is a gulf between an order and its execution. It has to be filled by the act of understanding". Wittgenstein's way to avoid the paradox is, according to Finkelstein, to show that the existence of a gulf is the result of a philosophical misconception. The misconception is a result of generalizing from those cases where someone misinterprets a sentence, or an order, or a signpost, to all cases. To view an order as nothing but sounds, or ink-marks, goes hand in hand with the

24 As Bar-On insists, semantic continuity neither implies semantic equivalence nor is inconsistent with the issue that some terms in avowals, such as 'I', refers differently from other ascriptions to the same subject.

idea that every possibility of misunderstanding a sentence, or a rule, should be eliminated. But, there does not seem to be any reason to view cases where understanding goes smoothly under this light. In the normal case we do catch the rule:

> A child might misunderstand the instruction "Beat six egg whites until stiff peaks form". […] It doesn't follow that I need an interpretation in order to understand these words when I encounter them in a cookbook. For me, there is no gulf between such an instruction and what it requires; I see what it calls for – without the need for interpretation or explanation. (Finkelstein, 2008: 81)

Following this line of reasoning, Finkelstein insists that there is no gap either between our avowals and the alleged mental states behind them. It makes perfect sense to view our self-ascriptions as being one of our ways to express our psychological states and it is equally plausible to claim that we are literally capable of perceiving the mental life of others in their behavior. This is very far from an image of behavior as mere movement divested of psychological import and always in need of interpretation. In this sense, avowals express mental states with truth value because, unlike a natural expression e.g. a smile, they also have an assertoric function. For Finkelstein this should explain first-person authority: usually my words and my face are the best place to start if you want to know about my mental life (Finkelstein, 2008: 100-01).

Bar-On also defends a version of expressivism –which she calls 'Neo-Expressivism'– that doesn't oppose expression to description, yet following a very different strategy and reasons from the ones we find in Finkelstein. While for Finkelstein an explanation of avowals should maintain "Intimacy", "Naturalness" and "Responsibility" as important traits of them, for Bar-On, an account of avowals should be able to respect two main elements: semantic continuity – "the claim that avowals are interchangeable *salva veritate* in context with certain unproblematic statements and can figure in certain logical inferences" (Bar-On, 2004: 10)– and epistemic asymmetry – "the

claim that there are genuine and important epistemic contrasts between avowals and their semantic cousins" (Bar-On, 2004: 10). She maintains the first factor by defending that an avowal expresses a subject's first-order condition *and* the subject's higher-order judgment that she is in that condition (Bar-On, 2004: 305). According to her, the avowing subject performs similar acts in natural expressions and avowals; however, the results differ in the sense that avowals, but not natural expressions, have semantic structure (Bar-On, 2004: 255)[25]. The second factor is explained with reference to an asymmetric presumption of truth in favor of avowals and by the "ascriptive immunity to error" –in addition to the immunity to error through misidentification– that avowals present.

In brief, Bar-On extends the phenomenon of "immunity to error through misidentification", such as defended by Shoemaker (1968: 82), Evans (1982: 218) and Strawson (1966: 165)[26], to the other components of avowals. If there is no sense for one that affirms to feel pain to say "there is someone that feels pain, but is it me?"

---

25 Another way of explaining this point is by making reference to different senses of 'expressing'. Bar-On finds in Sellars (1956) the following distinction: "EXP_1 the *action* sense: a *person* expresses a state of hers by intentionally doing something; EXP_2 the *causal* sense: an *utterance* or piece of behavior expresses an underlying state by being the culmination of a causal process beginning with that state; EXP_3 the *semantic* sense: e.g., a *sentence* expresses an abstract proposition, thought, or judgment by being a (conventional) representation of it" (Bar-On, 2004: 216). Contrary to a somehow traditional way of distinguishing between non-linguistic expressions and linguistic expressions in terms of EXP_2 and EXP_1 respectively, Bar-On defends that both can express in the sense of EXP_1 and EXP_2, but only linguistic expressions express in the sense of EXP_3.

26 Strawson (1966) doesn't use the term 'immunity to error through misidentification'. But, as pointed out by Evans (1982), he clearly refers to this same phenomenon. Strawson explains "When a man (a subject of experience) ascribes a current or directly remembered state of consciousness to himself, no use whatever of any criteria of personal identity is required to justify his use of the pronoun 'I' to refer to the subject of that experience. It would make no sense to think or say: *This* inner experience is occurring, but is it occurring to *me*? (This feeling is anger; but is it I who am feeling it?)" (Strawson, 1966: 165).

for Bar-On, such immunity applies to the entire avowal: there is no sense either in saying that "I feel something, but is it pain?" She explains:

> We saw that Shoemaker and Evans both offer the following intuitive test for immunity to error through misidentification [IETM]. When a self-ascription of the form "I am F" is IETM, then, although I may fail to be F, so my self-ascription may be false, there is no room for me to think: "*Someone* if F, but is it me?" […] This is because in such cases I have no grounds for thinking that someone has the relevant properties over and above, or separately from, any grounds I might have for thinking that I have them. […]
>
> Now consider the ascriptive part of avowals. In the normal case, as I say or think, "I am feeling terribly thirsty", it would seem as out of place to suggest, "I am feeling *something*, but is it thirst?" as it would to question whether it is I who am feeling the thirst. Or take an avowal with intentional content, such as "I'm really mad at you". "I am mad at *someone*, but is it *you*?" and "I'm in *some* state, but is it being mad?" would both be as odd as "*Someone* is mad at you, but is it I?" when I simply avow being mad at you (as opposed to making a conjecture about my own state of mind, for example). (Bar-On, 2004: 193)

However, the phenomenon Bar-On identifies as 'ascriptive immunity to error' doesn't seem to be as intuitive as she defends it to be. Although there seems to exist a sort of immunity the subject enjoys concerning self-identification, it is not so easy to see this immunity governing other components of avowals. In case I avow that "I feel terribly annoyed by the crowd", it seems perfectly fine to ask myself "I feel terribly annoyed, but is it by the crowd?" and indeed realize that "I feel terribly annoyed not by the crowd, but by not having had lunch yet". It seems that if expressionism is dependent on appealing to such immunity, it has less chances of being successful. In that sense, Finkelstein model seems to present an advantage over neo-expressivism insofar as it doesn't need to refer to such real epistemic advantages of avowals. According to his view, the security of avowals

is not a matter of being immune to errors, but a matter of expressing directly one's mental states. It is a more modest explanation, but it seems to be sufficient for explaining in part the specialness of first-person perspective.

3. The puzzle of explaining first-person authority through first-person perspective. Authority as an attribute of a person rather than of a particular self-ascription.

In the first section, I examined a sort of dilemma between two traditional paradigms on self-knowledge, the Cartesian and the Rylean approaches. I defended that neither of them were satisfactory insofar as they are compatible with the inexistence of one or another perspective on oneself. I argued that we are able to comprehend self-knowledge once we understand a person as being responsible for one's judgments and sensible to others' reasoning. I assumed the notion of third-person perspective to be out of dispute, contrary to what happens to first-person perspective. I dealt with that question in the second section. There, I examined three general modes of explanation on first-person perspective: detectivism –in its old and new versions-, constitutivism, and expressionism –also under two main variations of it.

Among other discrepant elements, those models involve the election among three paradigms, this time, of first-person perspective: report, constitution or expression. First, we were presented with the idea that thinking or speaking about one's mind from the first-person perspective consists of reporting mental states and contents. In a sense, the dilemma between Descartes and Ryle is part of such a paradigm. In one extreme, a subject reports inner items through acceding to them by inner perception, whereas at the other extreme, a person also reports one's mental states by acceding to one's behavior. I have argued against such a general paradigm,

examining three significant criticisms. The next move was to study alternatives to this highly problematic model: constitutivism and expressionism. According to constitutivism, talking and thinking about one's mind from first-person perspective establish one's very mental conditions. And according to expressivism, such a perspective consists of expressing one's mind, where detection has no place at all. In a sense, both models arise as a reaction to the detectivist ideal. Nevertheless, Wright's constitutivism has also received a somewhat important criticism concerning the imputation of excessive responsibility to the subject. On the side of expressivism, we have seen three variants of it: the simple expressivism and the models developed by Bar-On and Finkelstein. Among the available options, I have favored Finkelstein's model. Simple expressivism has failed in characterizing several of the roles avowals ordinarily have, for example, in inferential reasoning. On the other hand, Bar-On's model is committed to doubtful conditions, such as the ascriptive immunity to error. A common factor among all those paradigms, however, is that they all assume that explaining the specialness of the first-person perspective gives us the clues to explain why we are legitimately attributed with authority over our statements about ourselves.

As I stipulated initially, avowals are self-ascriptions of mental states that are products of first-person perspective, no matter how we characterize such a perspective. Avowals contrast with others' attributions of mental states to me *and* with ordinary self-ascriptions produced from third-person perspective. It is generally assumed that, as part of such a contrast, avowals have a sort of security that others' ascriptions *and* non-avowals self-ascriptions don't have. When I affirm, for example, that "I believe this summer is getting incredibly hot", generally there is no space for others to doubt whether I really believe that, or for others to ask for the basis of my judgment. In contrast, it is perfectly fine to ask how one knows that "Merce believes this summer is getting incredibly hot". This

security aspect of avowals has been characterized as indicative of first-person authority. Insofar as this authority is an attributed property, the discussed models intend to search for the basis of such an authority –or for its legitimacy– on avowals' characteristics. In a sense, attributing authority is to recognize those special aspects on one's words.

However, what shows that one's self-ascriptions are avowals? That is, what shows that they are products exclusively of first-person perspective? It seems that nothing does. Unless one previously assumes that everything one says about oneself is produced through first-person perspective –such as Cartesianism– it seems reasonable to think that nothing in one's speech itself makes such a mark explicit[27]. A statement such as "I'm so tired" could be either an avowal or the result of my quick recognition of my tired face in a mirror. But, according to the standard view, I will be legitimately attributed with authority only in the first case. Let's consider another example. Let's suppose a friend of mine asks about my wish to sell my car, after all, he has been listening to my plans for months. But now I answer "I don't really want to sell it". I may have said that as an immediate answer to his question and the case is that I've changed my mind. However, it could be as well the conclusion of my reasoning over my behavior: although I did advertise my car on newspapers, I didn't put a very attractive ad; I took weeks to answer to the possible purchasers; I keep saying how happy I am with its

27 One could also establish a weaker condition: that most of my self-ascriptions are products of first-person perspective. However, I am not sure whether there are good arguments for defending that. Moreover, apart from the problem of demarcation I'm pointing at, it seems to exist a sort of circularity in explaining authority in terms of avowals: i. Avowals are assumed to be authoritative; ii. They are assumed to be authoritative because doubts about avowals are out of place [an alternative to step (ii) would be: they are assumed to be authoritative because they are more secure, and they are more secure because they are not open to doubts]; iii. There is no place for doubts because they are avowals. Therefore, avowals exhibit authority because they are avowals.

good functioning and with its beautiful design. From such past experiences, I've concluded that I don't really wish to sell my car.

Our puzzle is the following: what is the connection between explaining first-person authority in terms of first-person perspective if any given example of authoritative statement can be a product either of first or third-person perspective? While we have two perspectives on ourselves –one is supposed to be authoritative while the other is not– we only have one statement, which normally receives authority. In principle, attributing or accepting authority over one's statements doesn't require knowing what kind of reasoning a subject has taken, nor would it be a reasonable requirement. But, if this is so, where is the alleged intrinsic connection between first-person authority and first-person perspective? The available models that explain our special proximity to our minds seem to let intact the gap between what such models explain and the basis for authority's attribution. According to constitutivism, for example, when I answer to my friend from first-person perspective that I don't want to sell my car, what I do is to constitute my very state of mind; to constitute the very truth conditions of my statement. In contrast, when I give the same answer from the third-person perspective, such as the inferential reasoning about my past behavior, I don't constitute anything. At most, I report a state of mind already constituted in other circumstances. But, who knows which perspective I've taken? Despite authority being considered an exclusive trait of the first situation, the second one doesn't present anything that could prevent our interlocutor from also attributing authority to my words. Should we consider that our interlocutor incorrectly attributes authority to me in the second situation?[28]

28 Someone could answer yes and claim that the second situation is one of insincerity. In other words, not making explicit that one is taking a third-person perspective should amount to insincerity. However insofar as a person has two perspectives on oneself, she doesn't stop being herself if one perspective precedes the other in particular cases. More than that, it seems indeed that there are

The given examples don't seem to suggest that I change from a situation where I don't have authority –e.g. when I say "I'm terribly tired" by recognizing me in the mirror– to another situation where I do possess authority –e.g. when I spontaneously say that "I'm terribly tired". In the first situation, my interlocutor would unlikely reply "How are you so sure? You just looked at the mirror!" Depending on the situation, she will instead take my self-ascription for granted and answer something like "I'm so sorry, why don't you take a rest?"; exactly the same answer one would probably give to a proper avowal. After all, I don't cease to be me when I infer from or observe my own behavior. In that sense, it seems that I don't lose my authority when I engage in third-person perspective and our interlocutor doesn't err in attributing it to me in such circumstances. This is indicative of the fact that more than recognizing a statement as authoritative, what we recognize is the very person as an authority over her statements, no matter the course of reasoning she has taken. This is a consequence of what I have been stressing, that both first and third-person perspectives on oneself are necessary for one to have self-knowledge, since the very revision of one's thoughts involves both perspectives. First-person perspective is still crucial, and giving a proper account of it is still a challenging inquiry. However, it is not mandatory to base one's authority on it. Since first-person authority concerns authority related to self-knowledge, and since two perspectives are necessary for having self-knowledge, authority may find its legitimacy in the interaction of those perspectives.

Richard Moran (2001) –who maintains first-person authority as still subsidiary of first-person perspective– has interestingly argued for the difference and the dynamic relation between the two kinds

cases where it isn't clear what perspective one is taking on oneself and hence, no principled way to separate authoritative from non-authoritative situations on the lines of a first/third person division. Asking for an additional mental state that evaluates which perspective one is taking seems to impute an artificial and useless condition for assuring first-person authority.

of perspectives. In his terminology, we have two kinds of attitudes on ourselves: a theoretical and a deliberative one, which correspond respectively to the third and first-person perspectives. The answer to theoretical questions involve the discovery of previously unknown facts about oneself, while the response to a practical question is reached through some decision or commitment which does not imply previous ignorance of facts about oneself (see Moran, 2001: 58). More precisely, adopting a deliberative attitude towards oneself leads to the acquisition of new beliefs or desires in view of the clarification of one's actual cognitive structure. While the theoretical attitude brings to light previously ignored but existent beliefs and desires, the deliberative attitude reasons about what to belief or desire focusing on what is the case and what is worth wanting. This sort of question, which terminates in the formation or endorsement of an attitude, contrasts with another class of inquiry (the theoretical one) which terminates in a description of one's states (Moran, 2001: 64). That is why, according to Moran, first-person perspective involves a good degree of responsibility: "the special responsibilities the person has in virtue of the mental life in question being *his own*" (Moran, 2001: 32). Moran insists that such differences are graphic once we introduce the question of transparency of first-person self-ascriptions. And Moorean sentences, such as "I believe it's raining but it is not raining", provide a useful context to such a discussion:

> For empirically, I can well imagine the accumulated evidence suggesting both that I believe that it's raining, and that it is not in fact raining. Theoretically, these are perfectly independent matters of fact, and I can in principle recognize the possibility of their co-occurrence, just as I can imagine my future conduct clashing with what I now decide to do. But, as I conceive of myself as a rational agent, my awareness of my belief is awareness of my commitment to its truth, a commitment to something that transcends any description of my psychological state. And the expression of this commitment lies in the fact that my reports on my belief are obliged to conform to the condition of transparency: that I can report on my *belief* about X

by considering (nothing but) X itself. (Moran, 2001: 84)

According to Moran, the important contrast between perspectives is the one between discovery and decision, where first-person perspective is characterized by the latter. And because first-person perspective is partly characterized by the responsibility one has in assuming one's beliefs as being one's own, first-person authority should be considered more as a demand than a concession –or attribution– from others to me. In this account, possessing first-person perspective is therefore a "normative demand". After all, affirming a Moorean sentence, for example, is an irrational behavior to take (as we have defended in the previous chapter). It seems indeed that in case one doesn't assume one's beliefs as being one's own, such a person is very likely to escape from the rational realm. This seems to favor Moran's account. However, as far as I can see, under this approach first-person perspective remains a demand on rationality, not on authority[29].

In the last section we have seen three paradigms of explaining first-person perspective: discovery, constitution (or decision) and expression. We have largely argued against the paradigm of discovery due to several reasons. In addition, as pointed out by Moran, such a paradigm in its new version seems to be better for characterizing a third-person perspective. Our range of choices was therefore reduced to constitutivism and to expressivism. Moran's account, which seems to favor constitutivism since he favors the paradigm of decision (although in a very different manner from Crispin Wright)[30] reveals

29 Moran argues that "The dimension of endorsement is what expresses itself in one aspect of first-person authority, where it concerns the authority of the person to make up his mind, change his mind, endorse some attitude or disavow it. This is a form of authority tied to the presuppositions of rational agency and is different in kind from the more purely epistemic authority that may attach to the special immediacy of the person's access to his mental life" (Moran, 2001: 92).

30 Apparently, Moran doesn't defend that avowals establish the very truth conditions of mental states.

the aspect of responsibility involved in avowals: the responsibility of self-ascribing mental states as being one's own, which exhibits transparency. However, we have seen another important element brought by expressivism: there is a relevant similarity between my non-linguistic expressions and my avowals. One can, most of the time, rely on both in order to know or describe one's mental states: although an avowal such as "I'm so happy to see you again" has a semantic structure, it seems to express the same mental state of happiness that my smile could express.

Finkelstein emphasizes Wittgenstein's idea that, as we learn how to speak, we acquire the capacity to replace, say, crying behavior for verbal expressions of pain and, latter on, we learn to think, even without any verbal expression, "That hurts" (Finkelstein, 2008: 112). Those two elements, however, are not incompatible. We can consider avowals, e.g. "I believe today will be a very hot day" as both expressing my mental state and exhibiting transparency. For when I state such a thing, I'm probably thinking about the weather. And I do express my belief that today will be a very hot day. My mental states have to do with how I see and interact with the world; and expressing them either by linguistic and non-linguistic behavior clearly shows how I think; it shows, for example, what I take to be true or false[31]. This is what I express by avowals.

This is the specialness about first-person perspective. The characteristics usually attributed to avowals such as groundless, immediacy and commitment to its truth, are then comfortable maintained by this sort of hybrid model of first-person perspective. However, things are not so easy with respect to the characteristic of being authoritative. Even if we consider that expressions are more secure than inferred self-ascriptions, we still have to deal with the

31 In Borgoni (2008, section 3) I argue for this point. A person's statement conveys simultaneously information about the matter in question and information about one's mind.

proposed puzzle. But the solution I'm insisting on is that avowals don't need to support one's authority by themselves. Nevertheless, we could still insist on the question whether avowals are more secure or not than third-person self-ascriptions.

Avowals are usually supposed to exhibit a presumption of truth in their favor, and also to exhibit a presumption that the "avower" knows what is saying. I am considering avowals as being expressions and as such, they are reliable bases for the apprehension of one's mental states. Insofar as I am who expresses my mental states, I can be said to know them. However, considering avowals as expressions doesn't ban the possibility that a certain avowal expresses the wrong mental state, such as the sleepy child who cries asking for playing, when she is suffering because what she *really* wants is going to bed. There is a sense in which we learn to express our mental states accurately, which seems to be a basic requirement of an agent who reasons (indeed transparency seems to have an important role here). Therefore, theoretically, avowals seem to exhibit the sort of security –although not an infallible one– necessary to explain the attribution of truth and authority to one's self-ascriptions. However, as I have defended, such an attribution applies to a broader phenomenon than only to avowals; it applies to almost all self-ascription one states. I defend that part of the explanation about authority lies on some pragmatic traits of our communicative practices, and part on the epistemic advantage the subject has by being able to contrast one's perspectives.

Let's return to the presumption of truth, which is supposed to be an exclusive attribute of avowals. First, the puzzle seems to suggest that not only avowals are supposed to be true. It seems indeed that most of one's mental self-ascriptions enjoy such an attribute. Second, not only self-ascriptions, avowals or not, are taken for granted. In a conversation, much of what is said enjoys this characteristic. That is, part of the presupposition of truth in one's mental self-ascription seems to be located in a more general trait of our communicative

interactions. Understanding what one says requires from me to keep lots of things out of doubt. I just cannot follow a conversation while doubting or asking for reasons of every sentence my interlocutor affirms. This general pragmatic trait should partly explain the usual presumption of truth one's self-ascriptions has. However, first-person authority clearly counts for more.

As I have defended, first-person authority is *attributed* and *alienable*. I'm attributed by my fellows with authority over my mind, and once I have it, it is not easily detached from me. Although the subject has a perspective on herself which only she has, her real epistemic advantage lies on being able to engage both in first and third-person perspectives. She is able, for example, to perceive that one's avowal is in conflict with another and then correct them. In that sense, revising one's own thoughts is the person's business. Following Burge's (1998b: 258) argumentation, one of the particularities of first-person perspective is revealed by the fact that reasoning from such a perspective involves immediate implementation of the evaluation on the evaluated attitudes. However, for something to count as a reason to maintain or to change my own judgment about my states of mind, it will probably require some extent of third-person perspective: it will require, for example, following someone else's reasoning about my behavior or following an inferential path from self-analysis. One important characteristic of self-ascriptions is that they usually don't need to present any reason in their support. This is part of the very phenomenon of first-person authority. However, such characteristic doesn't avoid the occurrence, for example, of cases where there are conflicting evidences, such as when one's behavior contrasts with one's self-ascription. In such circumstances, one needs to be able to reason about one's own self-ascriptions and, in a sense, one is ultimately responsible for correcting one's own thoughts. I have defended the necessity of third-person perspective for an account of self-knowledge, appealing to cases where someone ignores or is mistaken about some of one's mental states. In cases where

refocusing my attention is useless, I will not know anything new about my mind if I keep insisting on my first-person perspective. It seems that it is necessary to bring into *my* perspective pieces of information that are accessible from third-person perspective. Once such perspective provides me with information –reasons– in support of or against my self-judgments, they immediately incorporate my evaluation while I'm reasoning from first-person perspective. Again, first-person perspective has several particularities as compared to the third-one: avowals express one's thoughts, exhibit transparency, and exhibit the immediate incorporation of the result of reasoning. But it is in cooperation with third-person perspective that first-person perspective allows one to be taken as an authority over one's thoughts.

My proposal is, therefore, not a proper epistemic approach, since I don't defend we possess some kind of special knowledge of our own minds in terms of the epistemic qualities of first-person mode of knowledge[32]. However, I do argue for an epistemic advantage of the subject which gives legitimacy for our authority over our minds. This model of authority can be seen as part of what Bar-On identifies as the 'model of expertise'. According to her, under such a model – which she criticizes[33]– "people who have authority are individuals whom we take to have greater knowledge than most of us about certain matters, through having greater experience, training, or

32 I'm considering that the phenomenon of immunity to error through misidentification cannot provide the epistemic basis for a real epistemic advantage of first-person perspective.

33 So, contrary to Dorit Bar-On, I'm defending that first-person authority is partly a matter of an epistemic advantage. She indeed prefers to use the term "first-person privilege" to replace the label 'first-person authority'. She says: "since authority is often understood in epistemic terms, in Bar-On and Long (2001), we introduced the notion of "first-person privilege" as a more neutral notion, which does not prejudge the question whether avowals security is a matter of epistemic advantage" (Bar-On, 2004: 123). Bar-On's worry concerns the location of epistemic advantage in the first-person perspective. Since I locate such an epistemic advantage in the dynamic between the two perspectives, I'm immune to her criticisms.

natural facility" (Bar-On, 2004: 123-24). She completes, "Experts are people who 'know best'". In the model of expertise, however, in spite of experts being considered to know best, their judgments are not guaranteed by their own authority. Authority has a legitimate basis, but it doesn't provide to any particular judgment a direct condition of being true. We could say that the same occurs in first-person authority. While we are legitimately attributed with first-person authority, because we can know our minds better since we can contrast our two perspectives, it is not the case that everything we self-attribute to ourselves is accurate. Others can indeed know certain aspects of our mind better than us under certain circumstances.

In brief, my proposal combines a non-epistemic approach to first-person perspective with an epistemic explanation of first-person authority. On the one hand, I defend a hybrid model of expressivism that incorporates transparency, and consequently incorporates responsibility. It could be seen indeed as a modified version of Finkelstein's model. First-person perspective's self-ascriptions present some security insofar as any expression presents it, but they are not products of any epistemic access to one's mind. Nor do they present epistemic traits such as ascriptive immunity to error. On the other hand, I have defended that the root of first-person authority is in fact epistemic; it lies on the range of possibilities the subject has to know one's thoughts.

Bibliographical References

Alston, W. 1971, "Varieties of Privileged Access", *American Philosophical Quarterly* 8, pp. 223-41.

Armstrong, D. M. 1968, "Introspection", in Q. Cassam (ed.) 1994, *Self-knowledge*, Oxford, Oxford University Press, pp. 109-17.

Bar-On, D. 2004, *Speaking My Mind: Expression and Self-Knowledge*, Oxford, Clarendon Press.

Bilgrami, A. 2006, *Self-Knowledge and Resentment*, Cambridge, Mass.,

Harvard University Press.
Borgoni, C. 2008, "Interpretando la Paradoja de Moore: la irracionalidad de una oración mooreana", *Theoria* 23/2 (62) pp. 145-61.
Borgoni, C. 2009a, "When Externalism and Privileged Self-knowledge are Compatible and When They are Not", *Episteme NS* 29 (1), forthcoming.
Burge, T. 1998b, "Reason and the First Person", in C. Wright, B. Smith & C. MacDonald (ed.) 1998, *Knowing Our Own Minds*, Oxford, Clarendon Press, pp. 243-70.
Davidson, D. 1984, "First Person Authority", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 3-14.
Davidson, D. 1987, "Knowing One's Own Mind", in D. Davidson 2001, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press, pp. 15-38.
Evans, G. 1982, *The Varieties of Reference*, New York, Oxford University Press.
Finkelstein, D. H. 2008, *Expression and the Inner*, Cambridge, Mass., Harvard University Press.
Moran, R. 2001, *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton, Princeton University Press.
Ryle, G. 1949, *The Concept of Mind*, London, Hutchinson's University Library.
Sellars, W. 1969, "Language as Thought and as Communication", *Philosophy and Phenomenological Research* 29 (4), pp. 506-27.
Shoemaker, S. 1968, "Self-Reference and Self-Awareness", in Q. Cassam (ed.) 1994, *Self-knowledge*, Oxford, Oxford University Press, pp. 80-93.
Shoemaker, S. 1994, "Self-Knowledge and 'Inner Sense': Lecture II: The Broad Perceptual Model", *Philosophy and Phenomenological Research* 54 (2), pp. 271-90.
Strawson, P. F. 1966, *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*, New York, Methuen & Co.
Williamson, T. 2000, *Knowledge and its Limits*, Oxford, Oxford University Press.
Wittgenstein, L. 1953, *Philosophical Investigations* (translated by G. E. M. Anscombe), Oxford, Basil Blackwell, 1958.
Wittgenstein, L. *Zettel* (translated by G. E. M. Anscombe), Oxford, Basil

Blackwell, 1967.

Wright, C. 1989 "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention", in B. Gertler (ed) 2003, *Privileged Access: Philosophical Accounts of Self-Knowledge*, Aldershot, Asghate. pp. 147-57.

Wright, C. 1998, "Self-Knowledge: the Wittgensteinian Legacy", in C. Wright, B. Smith & C. MacDonald (ed.) 1998, *Knowing Our Own Minds*, Oxford, Clarendon Press, pp. 13-45.

Wright, C. 2001, "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics", in C. Wright 2001, *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations*, Cambridge, Mass., Harvard University Press, pp. 170-213.

CONCLUSION

Some find it very puzzling to understand how one can know at first hand one's own thoughts if such thoughts are themselves constituted by external factors to oneself. I also find it puzzling, but only as long as we consider the mind as a self-illuminated entity, on the one hand, and if the external conditions of individuation correspond to atomic causal relations between entities in the world and entities in one's mind, on the other hand. That was the intermediate step taken in this dissertation; indeed, a central one. The fourth chapter, which dealt with the incompatibilist debate between self-knowledge and externalism, outlined one of the important conclusions driven by this dissertation: incompatibilism does hold, but in very particular situations; for a relevant variety of externalist accounts and conceptions of self-knowledge, they are compatible. *Reductio ad absurdum* arguments show Putnam's externalism to be incompatible with self-knowledge acquired from first-person perspective. And the old detectivist model of self-knowledge is incompatible with an externalist conception of the mind, under the conditions designed by slow-switching cases. But, in principle, one need not choose between holding externalism and believing that we have some extent

of privileged self-knowledge: other models of externalism and of self-knowledge coexist pacifically. However, my primary objective with this dissertation was to go beyond pacific coexistence and to promote integration. After all, the question about compatibility doesn't exhaust the possibilities of reasoning about self-knowledge within an externalist outlook of the mind.

The perplexity of finding out that one's own thoughts in fact don't depend only on oneself is partly dissolved through making explicit other forms of externalism and alternative accounts of self-knowledge. A relevant part of the first and of the last chapter explored, respectively, the plurality of these positions. The scope of externalist positions I studied includes Tyler Burge, Donald Davidson, Ludwig Wittgenstein, Timothy Williamson and John McDowell. And one of the parallel results reached by the first chapter was the classification of such positions into a matrix articulated in terms of the distinction between global and two-factor externalisms and of the contrast between extrinsic and constitutive externalisms. This classification provided a new way of charting the externalist landscape. It permitted, for example, to relocate the difference between Putnam and Burge from the physical / social disjunction to the two-factor / global distinction. Such a matrix also provided new tools for understanding some of the important contrasts among other externalist positions. However, the main result achieved in the first chapter was the defense of a specific sort of externalism. Such a position holds that the entire mind –its mental states and contents– is partly individuated by external factors to one's skin. The contrast between narrow and broad contents was consequently discarded, but this doesn't mean that the subjective realm was itself dissolved. The externalism I favored in the first chapter –and that I sustain along this dissertation– holds that it is possible to make room for subjectivity under global externalism. I take the notion of subjectivity to be closely related to the notion of self-knowledge acquired from first-person perspective. For this reason, the final result of keeping

subjectivity within a global externalism was only completed by the last chapter. Another important element of this externalist picture was the emphasis on the explanation of the externality of the mind in terms of the presence of knowledge: the mind is only conceivable as such once it is populated by knowledge. Extrinsic and constitutive externalisms correspond to two levels of explanation about the externality of the mind that are not incompatible. However, I defended the primacy of what I called 'constitutive externalism'. Appealing exclusively to causal relations between mind and world (as the extrinsic path does) only allows us to conceive mental states as a combination of an internal and of an external aspect to the mind. This makes it hard to understand how external factors to one's skin could really be part of one's mind.

The sort of externalism defended in this dissertation is different from Putnam's in various senses, including from the aspect that makes Putnam's externalism incompatible with self-knowledge. For this reason, my position overcomes the incompatibilist challenge designed in terms of *reductio ad absurdum* arguments. The overall position defended in this dissertation also responds to the incompatibilist challenge designed in terms of slow-switching cases. The model of self-knowledge defended in the sixth chapter differs enormously from the old detectivist model. I defended that a model of self-knowledge should account for the double aspect of knowing one's own thoughts. I argued for the necessity of making room for first and third-person perspectives in order to understand how one revises, and consequently, how one knows one's own thoughts. Therefore, one relevant result reached in the last chapter was the defense of an account of self-knowledge composed by these two perspectives. The account of self-knowledge acquired by first-person perspective I defended is a hybrid model of expressivism that incorporates transparency. Avowals –self-ascriptions of mental states that are products of first-person perspective– are relevantly similar to non-linguistic expressions insofar as they directly express one's

mental state. But, unlike non-linguistic expressions, avowals have a semantic structure. I defended that avowals, such as "I believe it is going to rain soon", express one's mental state (in this case, the mental state of believing that it is going to rain soon), express one's judgment about one's own mental state ("I believe it is going to rain soon") and exhibit transparency. In stating such a sentence, one is probably thinking in terms of first-order belief, i.e., one is thinking that it is going to rain. This amounts to a non-epistemic picture of first-person perspective. However, I considered self-knowledge and the related explanation of first-person authority as remaining epistemic in its roots. The explanation of first-person authority was the central result achieved in the last chapter. I argued that first-person authority should not be explained in terms of first-person perspective, but instead in terms of the person's epistemic advantage to know her own thoughts, which is composed by first-person perspective but also by the third-person one. For this reason, I insisted that first-person authority was a person's attribute instead of a derivative attribute of avowals.

Other parallel results achieved in the other chapters were:

i. Analysis of some arguments, all of them highly influential in contemporary philosophy, that motivate externalism, which were neither classical references within the externalist literature, nor based on thought experiments: Wittgenstein's private language argument and his rule following considerations, Quine's rejection of the two dogmas of empiricism, Davidson's rejection of the third dogma and McDowell's rejection of the highest common factor (Chapter 2). My position in the thesis was influenced by Wittgenstein (especially under an interpretation close to the one developed by Finkelstein) and by Davidson (that in a sense includes Quine's argument);

ii. Location of some of the externalist arguments present in Wittgenstein's, Davidson's and McDowell's positions, which are generally identified as being externalist, but it is not always straightforward which of their arguments could favor such a position

and in what sense they do it (Chapter 2);

iii. Identification of Davidson's externalism, from the analysis of several theses spread over his works. This analysis concluded that Davidson's position makes room for two different, potentially conflicting, senses of externalism (Chapter 3);

iv. Proposal of a solution to Moore's paradox that outlines the irrational aspect involved in stating or thinking a Moorean sentence (Chapter 5). The irrationality of a Moorean sentence was revealed by the fact that one is expected to present a sort of unity that is under threat in Moorean cases; a unity that is also threatened by conceptions of self-knowledge that take it to be achieved only by means of one kind of perspective towards oneself (Chapter 6)

I opened this work by saying that knowing the nature of one's mind doesn't provide self-knowledge in the ordinary sense. However, I believe that it does give some important clues for conceiving self-knowledge. This was the inspiration for this work. My defense of an expressivist account of self-knowledge acquired from first-person perspective, with the incorporation of transparency, depended on some issues that exceeded the scope of the externalist discussion. However, such an account had its basis on the general externalist picture of the mind I intended to encourage. Already in my defense of global externalism in the first chapter, my worry was to embrace a form of externalism that didn't need to sacrifice first-person perspective; even a robust version of externalism, where the world constitute the mind in all of its aspects, doesn't need to deprive the subject of her own mind. In such a context, transparency was a crucial element to highlight the idea that having first-person perspective should not be conceived as an inner process of inspection of one's thoughts, but rather as a perspective directed outwards. In the last chapter, I reinforced the idea that the person's authority over her mind needs not to be conceived as the product of a perspective that only the subject has, namely, the first-person one. Authority is rather a product of the integration of the two perspectives the

subject can articulate: first and third-person perspectives. In this sense, to be attributed with authority over one's own thoughts is still legitimated by the subject's special condition. But this condition is rather characterized as the possibility of integrating, among other things, other's observations into one's own picture.

BIBLIOGRAPHICAL REFERENCES

Acero, J. J. & Calvo, T. (eds.) 1987, *Symposium Quine*, Granada, Universidad de Granada.

Albritton, R. 1995, "Comments on 'Moore's Paradox and Self-knowledge'", *Philosophical Studies* 77 (2/3), pp. 229-39.

Alston, W. 1971, "Varieties of Privileged Access", *American Philosophical Quarterly* 8, pp. 223-41.

Armstrong, D. M. 1968, "Introspection", in Q. Cassam (ed.) 1994, pp. 109-17.

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

Bar-On, D. 2004, *Speaking My Mind: Expression and Self-Knowledge*, Oxford, Clarendon Press.

Beebee, H. 2001, "Transfer of Warrant, Begging the Question and Semantic Externalism", *The Philosophical Quarterly* 51 (204), pp. 356-74.

Bensusan, H. & Pinedo, M. 2007, "When my Own Beliefs are not First-personal Enough", *Theoria* 58, pp. 35-41.

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

Bishop, M. 1999, "Why Thought Experiments are Not Arguments", *Philosophy of Science* 66, pp. 534-41

Black, M. 1952, "Saying and Disbelieving", *Analysis* 13, pp. 25-33.

Boghossian, P. & Peacocke, C. (eds.) 2000, *New Essays on the A Priori*, Oxford, Oxford University Press.

Boghossian, P. 1989a, "Content and Self-Knowledge", in P. Ludlow & N. Martin (eds.) 1998, pp. 149-73.

Boghossian, P. 1989b, "The Rule-Following Considerations", *Mind* 98, pp. 507-49.

Boghossian, P. 1998, "What the Externalist Can Know 'A Priori'", *Philosophical Issues* 9, pp. 197-211.

Bokulich, A. 2001, "Rethinking Thought Experiments", *Perspectives on Science* 9 (3), pp. 285-307.

Borgoni, C & Palomo, J. 2006, "Humildad Davidsoniana: Conocimiento e Ignorancia como Precondiciones de la Interpretación Radical", in F. Martínez-Manrique & L. Peris-Viñé (eds.) 2006, *Actas del V Congreso de la SLMFC*, Granada, Ediciones Sider S.C., pp. 105-9.

Borgoni, C. & Pedroso, M. 2004, "Is Nagel Davidsonable?" in J. C. Marek & M. Reicher (ed.) 2004, *Experience and Analysis* (27th International Wittgenstein Symposium), Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society, pp. 52-4.

Borgoni, C. 2006, "Davidson on Intercultural Dialog", in Gasser *et al.* (eds) 2006, *Cultures: Conflict – Analysis - Dialog* (29th International Wittgenstein Symposium), Kirchberg am Wechsel, Austrian Ludwig Wittgenstein Society, pp. 47-9.

Borgoni, C. 2008, "Interpretando la Paradoja de Moore: la irracionalidad de una oración mooreana", *Theoria* 23/2 (62), pp. 145-61.

Borgoni, C. 2009a, "When Externalism and Privileged Self-knowledge are Compatible and When They are Not", *Episteme NS* 29 (1), forthcoming.

Borgoni, C. 2009b, "En casa, en el mundo: el externismo global constitutivo", *Teorema* 23 (3), forthcoming.

Brandom, R. 2000, *Articulating Reasons*, Cambridge, Mass., Harvard

University Press.
Brandom, R. (ed.) 2000b, *Rorty and his Critics*, Oxford, Blackwell Publishers.
Brendel, E. 2004, "Intuition Pumps and the Proper Use of Thought Experiments", *Dialectica* 58 (1), pp. 88-108.
Brown, J. R. 1991a, *The Laboratory of the Mind*, London/New York, Routledge.
Brown, J. R. 1991b, "Thought Experiments: A Platonic Account", in T. Horowitz & G. Massey (eds.) 1991.
Brueckner, A. 1997, "Externalism and Memory", in P. Ludlow & N. Martin (eds.) 1998, pp. 319-31.
Burge, T. 1979, "Individualism and the Mental", in P. Ludlow & N. Martin (eds.) 1998, pp. 21-83.
Burge, T. 1982, "Other Bodies", in T. Burge 2007, pp. 82-99.
Burge, T. 1986, "Cartesian Error and the Objectivity of Perception", in T. Burge 2007, pp. 192-207.
Burge, T. 1988, "Individualism and Self-Knowledge", *The Journal of Philosophy* 85 (11), pp. 649-63.
Burge, T. 1989, "Wherein is Language Social?" in T. Burge 2007, pp. 275-90.
Burge, T. 1998a, "Memory and Self-knowledge", in P. Ludlow & N. Martin (eds.) 1998, pp. 351-70.
Burge, T. 1998b, "Reason and the First Person", in C. Wright, B. Smith & C. MacDonald (ed.) 1998, pp. 243-70.
Burge, T. 2003a, "Some Reflections on Scepticism: Reply to Stroud", in M. Hahn & B. Ramberg (ed.) 2003, pp. 335-46.
Burge, T. 2003b, "Replies from Tyler Burge", in M. Frápolli & E. Romero (eds.) 2003, pp. 243-96.
Burge, T. 2003/2006, "Descartes on Anti-individualism", in T. Burge 2007, pp. 420-39.
Burge, T. 2006, "Postscript to 'Individualism and the Mental'", in T. Burge 2007, pp. 151-81.
Burge, T. 2007, *Foundations of Mind*, Oxford, Oxford University Press.

Cassam, Q. (ed.) 1994, *Self-knowledge*, Oxford, Oxford University Press.

Chalmers, D. J. 1995, "Facing Up to the Problem of Consciousness", *Journal of Consciousness Studies* 2 (3), pp. 200-19.

Churchland, P. 1979, *Scientific Realism and the Plasticity of Mind*, Cambridge, Cambridge University Press.

Corbí, J. E. 1998, "A Challenge to Boghossian's Incompatibilist Argument", *Philosophical Issues* 9, pp. 231-42.

Dalmiya, V. 1990, "Coherence, Truth and the 'Omniscient Interpreter'", *The Philosophical Quarterly* 40 (158), pp. 86-94.

Davidson, D. 1963, "Actions, Reasons, and Causes", in D. Davidson 1980, pp. 3-19.

Davidson, D. 1967, "Causal Relations", in D. Davidson 1980, pp. 149-62.

Davidson, D. 1969, "The Individuation of Events", in D. Davidson 1980, pp. 163-80.

Davidson, D. 1970, "Mental Events", in D. Davidson 1980, pp. 207-27.

Davidson, D. 1971, "Agency", in D. Davidson 1980, pp. 43-61.

Davidson, D. 1973, "Radical Interpretation", in D. Davidson 1984b, pp. 125-39.

Davidson, D. 1973b, "The Material Mind", in D. Davidson 1980, pp. 245-59.

Davidson, D. 1973c, "Freedom to Act", in D. Davidson 1980, pp. 63-81.

Davidson, D. 1974, "On the Very Idea of a Conceptual Scheme", in D. Davidson 1984b, pp. 183-98.

Davidson, D. 1974b, "Psychology as Philosophy", in D. Davidson 1980, pp. 229-39.

Davidson, D. 1977, "The Method of Truth in Metaphysics", in D. Davidson 1984b, pp. 199-214.

Davidson, D. 1980, *Essays on Actions and Events*, Clarendon Press, Oxford.

Davidson, D. 1982, "Paradoxes of Irrationality", in D. Davidson 2004, pp. 169-88.

Davidson, D. 1983, "A Coherence Theory of Truth and Knowledge", in D. Davidson 2001, pp. 137-53.

Davidson, D. 1984, "First Person Authority", in D. Davidson 2001, pp. 3-14.

Davidson, D. 1984b, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press.

Davidson, D. 1987, "Knowing One's Own Mind", in D. Davidson 2001, pp. 15-38.

Davidson, D. 1988a, "Reply to Burge", *The Journal of Philosophy* 85 (11), pp. 664-65.

Davidson, D. 1988b, "The Myth of the Subjective", in D. Davidson 2001, pp. 39-52.

Davidson, D. 1990, "Epistemology Externalized", in D. Davidson 2001, pp. 193-204.

Davidson, D. 1991, "Three Varieties of Knowledge", in D. Davidson 2001, pp. 205-20.

Davidson, D. 1995, "Can There Be a Science of Rationality?", in D. Davidson 2004, pp. 117-34.

Davidson, D. 1999a, "Reply to A. C. Genova: The Very Idea of Massive Truth", in L. Hahn (ed.) 1999, pp. 192-94.

Davidson, D. 1999b, "Reply to Dagfinn Føllesdal", in L. Hahn (ed.) 1999, pp. 729-32.

Davidson, D. 2001, "Externalisms", in P. Kotatko, P. Pagin & G. Segal (eds.) 2001, *Interpreting Davidson*, Stanford, CSLI Publications, pp. 1-16.

Davidson, D. 2001b, *Subjective, Intersubjective, Objective*, Oxford, Clarendon Press.

Davidson, D. 2003, "Quine's Externalism", *Grazer Philosophische Studien* 66, pp. 281-97.

Davidson, D. 2004, *Problems of Rationality*, Oxford, Clarendon Press.

Davies, M. 2000, "Externalism and Armchair Knowledge", in P. Boghossian & C. Peacocke (eds.) 2000, pp. 384-414.

Davies, M. 2003, "Externalism, Self-Knowledge and Transmission of Warrant", in M. Frápolli & E. Romero (eds.) 2003, pp. 105-30.

Dennett, D. C. 1976, "Conditions of Personhood", in D. Dennett 1981, pp. 267-285.

Dennett, D. C. 1979, "True Believers", in D. Dennett 1987, pp.13-35.

Dennett, D. C. 1981, *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, Mass., MIT press.

Dennett, D. C. 1987, *The Intentional Stance*, Cambridge, Mass., MIT Press.

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

Finkelstein, D. H. 2008, *Expression and the Inner*, Cambridge, Mass., Harvard University Press.

Fodor, J. & LePore, E. 1992, *Holism: A Shopper's Guide*, Cambridge, Mass., Blackwell Publishers.

Foley, R. & Fumerton, R. 1985, "Davidson's Theism?" *Philosophical Studies* 48, pp. 83-9.

Frápolli, M. J. & Romero, E. (eds.) 2003, *Meaning, Basic Self-knowledge, and Mind*, Stanford, CSLI Publications.

Gendler, T. 2000, *Thought Experiment: On the Powers and Limits of Imaginary Cases*, NY, Routledge.

Gendler, T. 2005, "Thought Experiments", in D. Borchert (ed.) 2006 *Encyclopedia of Philosophy* 9, Detroit, Macmillan Reference, pp. 388-94.

Gert, B. 1986, "Wittgenstein's Private Language Argument", *Synthese* 68, pp. 409-39.

Gertler, B. (ed.) 2003, *Privileged Access: Philosophical Accounts of Self-Knowledge*, Aldershot, Asghate.

Gibson Jr., R. F. 2006, "Quine's Behaviorism cum Empiricism" in

R. Gibson (ed.) 2006b, pp.181-199.

Gibson Jr., R. F. (ed.) 2006b, *The Cambridge Companion to Quine*, Cambridge, Cambridge University Press.

Goldberg, S. 2003, "On Our Alleged A Priori Knowledge that Water Exists", *Analysis* 63 (1), pp. 38-41.

Goldman, A. 1979, "What Is Justified Belief?", in H. Kornblith (ed). 1994, pp. 105-30.

Goldman, A. 1994, "Naturalistic Epistemology and Reliabilism", in P. French, T. Uehling, & H. Wettstein. (eds.). *Midwest Studies in Philosophy 19: Philosophical Naturalism*, Notre Dame, University of Notre Dame Press, pp. 301-20.

Grice, H. P & Strawson, P. F. 1956, "In Defense of a Dogma", *The Philosophical Review* 65 (2), pp. 141-58.

Guttenplan, S. (ed.) 1977, *Mind and Language*, Oxford: Clarendon Press.

Hacker, P. M. S. 1990, *Wittgenstein: Meaning and Mind*, Oxford, Basil Blackwell.

Hahn, L. E. (ed.) 1999, *The Philosophy of Donald Davidson*, Chicago, Open Court.

Hahn, M. & Ramberg, B. (eds.) 2003, *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, Cambridge, Mass., MIT Press.

Hahn, M. 2003, "When Swampmen Get Arthritis: 'Externalism' in Burge and Davidson", in M. Hahn & B. Ramberg (eds.) 2003, pp. 29-58.

Heal, J. 1994, "Moore's Paradox: A Wittgensteinian Approach", *Mind* 103, pp. 5-24.

Heil, J. 1988, "Privileged Access", *Mind* 97 (386), pp. 238-51.

Hintikka, J. 1962, *Knowledge and Belief*, Ithaca and London, Cornell University Press. [Reference found in the Summary of Chapter 5]

Hintikka, J. 1962, *Saber y Creer* (translated by J. J. Acero), Madrid, Editorial Tecnos, 1979. [Reference found in Chapter 5]

Horowitz, T. & Massey. G. (eds.), 1991, *Thought Experiments in Science and Philosophy*, Savage, MD, Rowman and Littlefield.

Hurley, S. 2006, "Varieties of Externalism" in R. Menary (ed), *The Extended Mind*, Ashgate, forthcoming.

Jackson, F. 1982, "Epiphenomenal Qualia", *Philosophical Quarterly* 32, pp. 127-36.

Jackson, F. 1986, "What Mary Didn't Know", *The Journal of Philosophy* 83 (5), pp. 291-95.

Kornblith, H. (ed.) 1994, *Naturalizing Epistemology*, Cambridge, Mass., The MIT Press.

Kotatko, P., Pagin, P. & Segal, G. (eds.) 2001, *Interpreting Davidson*, Stanford, CSLI Publications.

Kriegel, U. 2004, "Moore's Paradox and the Structure of Conscious Belief", *Erkenntnis* 61, pp. 99-121.

Kripke, S. 1982, *Wittgenstein on Rules and Private Language*, Oxford, Basil Blackwell.

Kuhn, T. 1964, "A Function for Thought Experiments", in. T. Kuhn 1977, pp. 240-265.

Kuhn, T. 1977, *The Essential Tension*, Chicago, University of Chicago Press.

Lafont, C. 2005, "Was Heidegger an Externalist?", *Inquiry* 48 (6), pp. 507-32.

Lee, B. D. 2001, "Moore's Paradox and Self-Ascribed Belief", *Erkenntnis* 55, pp. 359-70.

Lewis, D. 1996, "Elusive Knowledge", *Australasian Journal of Philosophy* 74, pp. 549-67.

Linville, L. & Ring, M. 1991, "Moore's Paradox Revisited", *Synthese* 87, pp. 295-309.

Ludlow, P. & Martin, N. (eds.) 1998, *Externalism and Self-knowledge*, Stanford, CSLI Publications.

Ludlow, P. 1995a, "Externalism, Self-Knowledge, and the Prevalence of Slow Switching", in P. Ludlow & N. Martin (eds.) 1998, pp. 225-30.

Ludlow, P. 1995b, "Social Externalism, Self-Knowledge, and Memory", in P. Ludlow & N. Martin (eds.) 1998, pp. 307-10.

Macarthur, D. 2003, "McDowell, Scepticism, and the 'Veil of Perception'", *Australasian Journal of Philosophy* 81 (2), pp. 175-90.

Manning, R. N. 1995, "Interpreting Davidson's Omniscient Interpreter", *Canadian Journal of Philosophy* 25 (3), pp. 335-74.

Marton, P. 1999, "Ordinary versus Super-omniscient Interpreters", *The Philosophical Quarterly* 49 (194), pp. 72-7.

Marvan, T. (ed.) 2006, *What Determines Content? The Internalism / Externalism Dispute*, Cambridge, Cambridge Scholars Press.

McCulloch, G. 2003, *The Life of the Mind: an essay on phenomenological externalism,* London, Routledge.

McDowell, J. 1977, "On the Sense and Reference of a Proper Name", *Mind* 86, pp. 159-85.

McDowell, J. 1982, "Criteria, Defeasibility, and Knowledge", in J. McDowell 1998b, pp. 369-94.

McDowell, J. 1984a, "Wittgenstein on Following a Rule", in A. Miller & C. Wright (eds.) 2002, pp. 45-80.

McDowell, J. 1984b, "*De Re* Senses", in J. McDowell 1998b, pp. 214-27.

McDowell, J. 1992, "Putnam on Mind and Meaning", in J. McDowell 1998b, pp. 275-91.

McDowell, J. 1994, *Mind and World*, Cambridge, Harvard University Press.

McDowell, J. 1998, "One Strand in the Private Language Argument", in J. McDowell 1998c, pp. 279-96.

McDowell, J. 1998b, *Meaning, Knowledge & Reality*, Cambridge, Harvard University Press.

McDowell, J. 1998c, *Mind, Value and Reality*, Cambridge, Mass., Harvard University Press.

McGinn, C. 1982, "The Structure of Content" in A. Woodfield (ed.) 1982, pp. 207-58.

McKinsey, M. 1991, "Anti-Individualism and Privileged Access",

Analysis 51, pp. 9-16.

McKinsey, M. 2002, "Forms of Externalism and Privileged Access", *Philosophical Perspectives* 16, pp. 199-224.

McLaughlin, B. & Cohen, J. (eds) 2007, *Contemporary Debates in the Philosophy of Mind*, Oxford, Blackwell Publishers.

Miller, A. & Wright, C. (eds.) 2002, *Rule Following & Meaning*, Chesham, Acumen.

Miller, A. 2002, "Introduction", in A. Miller & C. Wright (eds.) 2002, pp. 1-15.

Moore, G. E. 1939, "Proof of an External World", *Proceedings of the British Academy* 25, pp. 273-300.

Moore, G. E. 1942, "A Reply to My Critics" in P. Schilpp (ed.) 1942, pp. 535-677.

Moore, G. E. 1944, "Russell's 'Theory of Descriptions'" in P. Schilpp (ed.) 1944, pp. 175-226.

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

Moya, C. 1998, "Boghossian's 'Reductio' of Compatibilistm", *Philosophical Issues* 9, pp. 243-51.

Nagel, T. 1974, "What is it Like to be a Bat?", *The Philosophical Review* 83 (4), pp. 435-50.

Norton, J. 1996, "Are Thought Experiments Just What You Always Thought?" *Canadian Journal of Philosophy* 26, pp. 333-66.

Norton, J. 2004, "On Thought Experiments: Is There More to the Argument?" *Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association, Philosophy of Science* 71, pp. 1139-51.

Nuccetelli, S. (ed.) 2003, *New Essays on Semantic Externalism and Self-Knowledge*, Cambridge, Mass., MIT Press.

Pessin, A. & Goldberg. S. (ed) 1996, *The Twin Earth Chronicles: twenty years of reflection on Hilary Putnam's "The meaning of 'meaning'"*, New York, Sharpe.

Pinedo, M. 2004, "The Anomalous Character of Experience",

in J. Marek & M. Reicher (ed.) 2004, *Experience and Analysis* (Proceedings of the 27th International Wittgenstein Symposium), Kirchberg am Wechsel (Austria): Austrian Ludwig Wittgenstein Society, pp. 269-71.

Pinedo, M. 2006, "Anomalous Monism: Oscillating between Dogmas", *Synthese* 148, pp. 79-97.

Prades, J. L. 2006 "Varieties of Internal Relations: Intention, Expression and Norms, *Teorema* 25 (1), pp. 137-54.

Preti, C. 1995, "Externalism and Analyticity", *Philosophical Studies* 79 (3), pp. 213-36.

Preti, C. 2002, "Normativity and Meaning: Kripke's Skeptical Paradox Reconsidered", *The Philosophical Forum* 33 (1), pp. 39-62.

Putnam, H. 1962, "It Ain't Necessarily So", *The Journal of Philosophy* 59 (22), pp. 658-71.

Putnam, H. 1970, "Is semantics possible?", in H. Putnam 1975b, pp. 139-52.

Putnam, H. 1973, "Explanation and reference", in H. Putnam 1975b, pp. 196-214.

Putnam, H. 1975, "The Meaning of 'Meaning", in H. Putnam 1975b, pp. 215-71.

Putnam, H. 1975b, *Mind, Language and Reality,* Philosophical Papers, vol. 2, Cambridge, Cambridge University Press.

Putnam, H. 1988, *Representation and Reality*, Cambridge, Mass., MIT Press.

Putnam, H. 1996, "Introduction", in A. Pessin & S. Goldberg (ed) 1996, pp. xiv-xxii.

Quesada, D. 1987, "Creencia, conducta y contexto", in J. Acero & T. Calvo (eds.) 1987, pp. 157-74.

Quine, W. v. O. 1951, "Two Dogmas of Empiricism", *The Philosophical Review* 69, pp. 20-43. [Reference found in the Summary of Chapter 2]

Quine, W. v. O. 1951, "Dos Dogmas del empirismo" (translated by

M. Sacristán), in *Desde un Punto de Vista Lógico*, Barcelona, Paidós, 2002, pp. 61-91. [Reference found in Chapter 2]

Quine, W. v. O. 1960, *Word and Object*, Cambridge, Mass., MIT Press.

Quine, W. v. O. 1969, "Epistemology Naturalized", in Kornblith, H. (ed.) 1994, pp. 15-31.

Quine, W. v. O. 1969b, *Ontological Relativity and Other Essays,* New York, Columbia University Press.

Quine, W. v. O. 1975, "Mind and Verbal Dispositions", in S. Guttenplan (ed.) 1977, *Mind and Language*, Oxford: Clarendon Press, pp. 83-95.

Quine, W. v. O. 1985, "States of Mind", *The Journal of Philosophy* 82 (1), pp. 5-8.

Ramberg, B. 2000, "Post-ontological Philosophy of Mind: Rorty versus Davidson", in R. Brandom (ed.) 2000b, pp. 351-77.

Recanati, F. 1993, *Direct Reference: From Language to Thought,* Oxford, Basil Blackwell.

Rudd, A. 1997, "Two Types of Externalism", *The Philosophical-Quarterly* 47 (189), pp. 501-07.

Ryle, G. 1949, *The Concept of Mind*, London, Hutchinson's University Library.

Sawyer, S. & Majors, B. 2005 "The Epistemological Argument for Content Externalism", *Philosophical Perspectives* 19, pp. 257-80.

Sawyer, S. 1997, *Semantic Externalism and Self Knowledge: Privileged Access to the World*, Phd Thesis, King's College London.

Sawyer, S. 1998, "Privileged Access to the World", *Australasian Journal of Philosophy* 76 (4), pp. 523-33.

Sawyer, S. 1999, "An Externalist Account of Introspective Knowledge", *Pacific Philosophical Quarterly* 80 (4), pp. 358-78.

Sawyer, S. 2002, "In Defence of Burge's Thesis", *Philosophical Studies* 107, pp. 109-28.

Sawyer, S. 2006, "Externalism, Apriority and Transmission of

Warrant", in T. Marvan (ed.) 2006, pp. 142-53.
Sawyer, S. 2007, "There is No Viable Notion of Narrow Content", in B. McLaughlin & J. Cohen (eds) 2007, pp. 20-34.
Schilpp, P. (ed.) 1942, *The Philosophy of G. E. Moore*, La Salle, Open Court.
Schilpp, P. (ed.) 1944, *The Philosophy of Bertrand Russell*, Evanston, Northwestern University.
Sellars, W. 1956, *Empiricism and the Philosophy of Mind*, reprinted in 1997, Cambridge, Mass., Harvard University Press.
Sellars, W. 1969, "Language as Thought and as Communication", *Philosophy and Phenomenological Research* 29 (4), pp. 506-27.
Shoemaker, S. 1968, "Self-Reference and Self-Awareness", in Q. Cassam (ed.) 1994, pp. 80-93.
Shoemaker, S. 1994, "Self-Knowledge and 'Inner Sense': Lecture II: The Broad Perceptual Model", *Philosophy and Phenomenological Research* 54 (2), pp. 271-90.
Shoemaker, S. 1995, "Moore's Paradox and Self-Knowledge", in S. Shoemaker 1996, pp. 74-93.
Shoemaker, S. 1996, *The First-person Perspective and Other Essays,* Cambridge, Cambridge University Press.
Strawson, P. F. 1966, *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*, New York, Methuen & Co.
Vega, J. 2006, "Appearances and Disjunctions: Empirical Authority in McDowell's Space of Reasons", *Teorema* 25 (1), pp. 63-81.
Warfield, T. 1992, "Privileged Self-Knowledge and Externalism are Compatible", in P. Ludlow & N. Martin (eds.) 1998, pp. 215-21.
Warfield, T. 1997, "Externalism, Privileged Self-knowledge, and the Irrelevance of Slow Switching", in P. Ludlow & N. Martin (eds.) 1998, pp. 231-37.
Warfield, T. 1998, "A Priori Knowledge of the World: knowing the world by knowing our minds", *Philosophical Studies* 92, pp. 127-47.

Williams, J. N. 1979, "Moore's Paradox – One or Two?", *Analysis* 39, pp. 141-42.

Williams, J. N. 2006, "Wittgenstein, Moorean Absurdity and its Disappearance from Speech", *Synthese* 149, pp. 225-54. Williamson, T. 2000, *Knowledge and its Limits*, Oxford, Oxford University Press.

Wittgenstein, L. 1953, *Investigaciones Filosóficas* (translated by A. García Suárez & U. Moulines), Barcelona, Crítica, 1988. [Reference found in Chapter 1, 2 and 5]

Wittgenstein, L. 1953, *Philosophical Investigations* (translated by G. E. M. Anscombe), Oxford, Basil Blackwell, 1958. [Reference found in the summaries of Chapter 1 and 2, in the translation of Chapter 5, and in Chapter 6]

Wittgenstein, L. 1969, *On Certainty* (translated by G. E. M. Anscombe & D. Paul), Oxford, Basil Blackwell, 1979.

Wittgenstein, L. *Zettel* (translated by G. E. M. Anscombe), Oxford, Basil Blackwell, 1967.

Woodfield, A. (ed.) 1982, *Thought and Object*, Oxford, Clarendon Press.

Wright, C. 1989, "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention", in B. Gertler (ed) 2003, pp. 147-57.

Wright, C. 1998, "Self-Knowledge: the Wittgensteinian Legacy", in C. Wright, B. Smith & C. MacDonald (ed.) 1998, pp. 13-45.

Wright, C. 2000, "Cogency and Question-Begging: Some Reflections on McKinsey's Paradox and Putnam's Proof", *Philosophical Issues* 10, pp. 140-63.

Wright, C. 2001, "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics", in C. Wright 2001b, pp. 170-213.

Wright, C. 2001b, *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations* Cambridge, Mass., Harvard University Press.

Wright, C. 2003, "Some Reflections on the Acquisition of Warrant by Inference", in S. Nuccetelli (ed.) 2003, pp. 57-77.

Wright, C., Smith, B. C. & MacDonald, C. (ed.) 1998, *Knowing Our Own Minds*, Oxford, Clarendon Press.