MENTAL CONTENT AND MENTALISTIC CAUSAL EXPLANATION: A CASE AGAINST EXTERNALISM

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

IŞIK SARIHAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

DECEMBER 2010

Approval of the graduate school of informatics

____

Prof. Dr. Nazife Baykal
Director

I Certify that this thesis satisfies all the requirements as a thesis for the degree of master of science.

____

Prof. Dr. Deniz Zeyrek
Head of Department

This is to certify that we have read this thesis and in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____________________	__________________
Assoc. Prof. Erdinç Sayan	Dr. Ceyhan Temürcü
Co-Supervisor	Supervisor

Examining Committee Members

Assoc. Prof. Cem Bozşahin	(METU, COGS)	___________________
Dr. Ceyhan Temürcü	(METU, COGS)	___________________
Assoc. Prof. Erdinç Sayan	(METU, PHIL)	___________________
Assist. Prof. István Aranyosi	(Bilkent, PHIL)	___________________
Assist. Prof. Annette Hohenberger	(METU, PHIL)	___________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Işık Sarıhan

Signature: ________________

ABSTRACT

MENTAL CONTENT AND MENTALISTIC CAUSAL EXPLANATION:

A CASE AGAINST EXTERNALISM

Sarıhan, Işık

M. Sc. Department of Cognitive Science

Supervisor: Dr. Ceyhan Temürcü

Co-Supervisor: Assoc. Prof. Erdinç Sayan

December 2010, 128 pages

This thesis presents a defense of the view that externalism cannot be a theoretical basis of a mentalistic causal-explanatory science, even though such a theoretical basis is implicitly or explicitly adopted by many cognitive scientists. Externalism is a theory in philosophy of mind which states that mental properties are relations between the core realizers of an individual’s mental states (such as brain states) and certain things that exist outside those realizers (such as what the content of a mental state corresponds to in the actual world.) After clarifications regarding the term “externalism” and reviewing the history and the various forms of the externalist theory, it is argued that the properties offered by externalist theories as mental properties have no causal influence on behavior, and therefore cannot causally explain it. The argument is largely based on a method of comparing the causal powers of entities which are identical in all respects except their mental properties (as construed by externalism), and the conclusions are supported by metaphysical reflections on causation, dispositions, relational properties and historical properties. Objections to the defended view are considered and refuted. The thesis is written in the style of modern analytic philosophy.

Keywords: Externalism, Mental Causation, Mentalistic Explanation, Mental Content, Philosophy of Cognitive Science

ÖZ

ZİHİNSEL İÇERİK VE ZİHİNSELCİ NEDENSEL AÇIKLAMA:

DIŞSALCILIĞA KARŞI BİR ARGÜMAN

Sarıhan, Işık

Yüksek Lisans, Bilişsel Bilimler

Tez Yöneticisi: Dr. Ceyhan Temürcü

İkinci Yönetici: Doç. Dr. Erdinç Sayan

Aralık 2010, 128 sayfa

Bu tez zihin felsefesinde "dışsalcılık" adıyla bilinen kuramın, her ne kadar birçok bilişsel bilimci tarafından açık ya da örtük biçimde doğruluğu kabul edilse de, davranışı zihinselci bir biçimde açıklamaya çalışan bir bilim dalının kuramsal temeli olamayacağını savunmaktadır. Dışsalcılık, zihinsel niteliklerin bireyin zihinsel durumlarının temel gerçekleştiricileri (örneğin, beyin durumları) ve o gerçekleştiricilerin dışındaki bazı varlıklar (örneğin, zihin-dışı dünyada var olan ya da var olmuş, bir düşüncenin içeriğine karşılık gelen nesneler) arasındaki ilişkiler olduğunu iddia eder. “Dışsalcılık” terimiyle ilgili netleştirmelerden ve dışsalcı kuramın tarihi ve çeşitleri üzerinde durduktan sonra, dışsalcı kuramların zihinsel nitelikler olarak önerdiği niteliklerin davranışa her hangi bir etkisi olmadığı ve bu nedenle davranışı açıklayamayacağı savunulacaktır. Tezde sunulan argümanın temelini, bütün nitelikleri aynı ama (dışsalcılığın önerdiği şekliyle) zihinsel nitelikleri farklı olan varlıkların nedensel güçlerinin karşılaştırılmasını içeren bir yöntem oluşturmaktadır, ve varılan yargılar nedensellik, eğilimler, ilişkisel nitelikler ve tarihsel nitelikler gibi varlıksal kategorilerle ilgili kavramsal çözümlemelerle de desteklenmiştir. Savunulan görüşe karşı ortaya koyulmuş argümanlar değerlendirilmiş ve çürütülmüştür. Tez, çağdaş analitik felsefe tarzında yazılmıştır.

Anahtar Kelimeler: Dışsalcılık, Zihinsel Nedensellik, Anlakçı Açıklama, Zihinsel İçerik, Bilişsel Bilimler Felsefesi

Gamze'ye

ACKNOWLEDGEMENTS

This thesis traveled four universities. In the spring semester of the year 2009, it began as a term paper for a class in Middle East Technical University on philosophy of biology, as a brief criticism of the teleosemantic theory of mental content. Thanks to the instructor of the class, Ayhan Sol, for his helpful comments on the paper.

Spring turned into summer, and the aforementioned paper turned into a talk in the beautiful resort town of Jurmala on the Baltic coast, in a workshop on mental causation, The First International Graduate School in Cognitive Sciences and Semantics. Thanks to those in the audience for their comments, particularly Jaegwon Kim, who made me aware of a serious objection to the view I have been defending, which I discuss in the thesis. Thanks to Sandra LaPointe and the University of Latvia for organizing the workshop, and to all the participants for the endless and enjoyable discussions over many cups of coffee, beer, and the Latvian sunset, which were not only instrumental in developing my views but also made me aware of the joys of being in a philosophical community.

The aforementioned talk turned into a paper written for the workshop. The paper, and this dissertation that grew from it, would be in a drastically

lower quality if it were not for the detailed and critical commentary of the paper's supervisor, Douglas Patterson.

Later on, the main ideas in the aforementioned paper were presented on Darwin Day in METU, the organizers of which I am grateful to, slightly before the ideas turned into a thesis proposal. I thank the METU cognitive science faculty for their commentaries during the presentation of the proposal.

Parts of the thesis were completed during an exchange term in the University of Osnabrück, during the spring term of 2010, and were presented in a departmental seminar. I thank the students and the faculty for their comments, particularly Sven Walter, Achim Stephan and Brian McLaughlin, comments which showed me that I still had some way to go to be persuasive. Another thing I learned during my time in Osnabrück was that some very special skills should be developed against people who don't believe that philosophy can accomplish anything in general. Thanks to fellow cognitive science student Chavdar Levkov for his total skepticism towards philosophy, even though he is not aware that that attitude itself is a philosophical one, and for the great time we have spent together. I should also mention that the local student community in Osnabrück is awesome, and I thank them for making my stay enjoyable, for showing up in my garden parties, and for all those tequila shots in Trash on Wednesday nights.

The thesis was completed in Budapest where I enroll as a PhD student in philosophy in Central European University as I write this preface. Some ideas about the definitional issues regarding externalism and internalism were presented in an in-house conference. Thanks to the CEU philosophy department for their feedback, particularly to my supervisor Katalin Farkas, who will be spending quite a lot of time in the forthcoming years discussing such issues with me.

Two philosopher friends shouldn't go unmentioned. Thanks to Pete Murray for lengthy e-mail exchanges about the problems related to the very idea of microphysical duplication of organisms, an idea the coherence of which is necessary for the thought experiments deployed in this thesis, and also for other online philosophical exchanges which happen every now and then. Thanks to Umut Baysan, again for many such exchanges, and for sharing with me the excitement of growing up to be an analytic philosopher of mind.

I also feel an obligation to thank my long-time band mates Özüm İtez, Hazar Mutgan and Gökçe Başar, even though it is not entirely clear how they relate to this thesis.

I am most grateful to the supervisors of this thesis, Ceyhan Temürcü and Erdinç Sayan, for their helpful commentaries and exchanges during the writing process. I hope they will also let me make this preface an occasion to apologize for the difficulties I created during the completion of the thesis, mostly due to my being continuously changing geographical locations.

Needless to say, without my dear family, this thesis wouldn't exist. This is not only because they are my biological creators, which is a trivial fact, and not only because they kept feeding me during all these years I chose to remain a full-time student. Their real contribution to the work is cultural: If this thesis succeeds in giving the impression of being written by an industrious person who can think clearly on a complicated subject and who is not reluctant to challenge the philosophical status quo, it is due to what the parents of the author passed on to him. If it doesn't succeed in such a thing, the parents are not to blame.

This thesis is written with the ultimate aim of contributing to our understanding of mentality and subjective experience, like any other thesis in philosophy of mind should be, and such understanding is valuable as far as it has a positive effect on our behavior towards those experiencing subjects, including our behavior in those very contexts where we are trying to acquire knowledge about the mind. In this thesis, I have deliberately tried to avoid referencing to books or articles that contain original experimental data acquired by tormenting and killing sentient non-human organisms.

However thankful I am to the people who have been around as I have been writing this thesis, the deepest gratitude and a humble immaterial gift goes to a person who hasn't been around during those two years, but for having already made my life a life worth living. It was a goodness which I did not fully deserve. I dedicate this thesis to her.

TABLE OF CONTENTS

CHAPTER 0

INTRODUCTION

Contemporary cognitive science is marked by at least two fundamental aims: The first one is explaining behavior through mental states and properties possessed by living things, or possibly, machines. Most or all such mental states are contentful states, states that are individuated by what one has in one's mind. The second aim is figuring out the place of such states and properties in the natural order. The first aim differentiates cognitive science from various methodological approaches such as behaviorism and mental eliminativism, approaches that are not committed to the existence of mentality at all. The second aim differentiates it from most mainstream psychology that is not necessarily interested in the natural foundations of mentality.

The second aim has an interesting relation to the first one, although it may not be so obvious at the first glance, a relation established by a worry had by many philosophers of cognitive science since the earliest days of the field. Many philosophers have been occupied with the worry that if we cannot identify mental properties with some properties in the so-called "natural" or "physical" world, then a mentalistic science of behavior is not possible. One reason behind this worry is a commitment to "naturalism" or "physicalism", the idea that everything that exists is physical, and if we cannot identify mental properties with properties that are known to the physical sciences, then mental properties do not exist.

Even if we leave this worry aside, there has been another one held by many philosophers: Behavior of intelligent animals, which is a collection of physical events, is sufficiently caused by other physical events in virtue of the physical properties of those events, and if mental properties are not strictly identical to these physical properties, then mental properties are *causally excluded* by physical ones.[1]

These worries have been among the important driving forces behind the recent philosophical attempts to find the place of mental properties in the natural world. How it is possible to find such an answer through philosophical methods is an interesting question, but it is beyond our current scope. Nevertheless, many of such attempts have fallen into a category known by the name of *externalism*. To put it roughly for now, a theory is externalistic if it states that having a mental state at a given time depends on certain relations between the mental vehicles of the entity which possesses the mental state and things outside those mental vehicles. A well known example is the causal-informational theory of mind, also held by many non-philosophizing practitioners in cognitive science, which holds

[1] This worry has been most famously voiced and explored through the years by Jaegwon Kim (1993, 2005), while Malcolm (1968) has been the earliest to draw attention to the problem.

that mentality is a matter of a body part's carrying causal information, although additional conditions might as well be necessary.

Externalist theories, depending on which theory we are talking about, claim that having a mind depends on an entity's having the following properties, which we will call "e-properties":

Synchronic-Relational: Being in a world[2] where there is X.

Historical-Causal: Being caused by X.

Historical-Use: Being a cause of X.

Normative: Having an etiological proper function.

This thesis is titled "Mental Content and Mentalistic Causal Explanation: A Case against Externalism." To paraphrase, it suggests that externalism has consequences such that if it is true, mental states do not have any causal efficacy or causal relevance in the world, at least not *qua* mental states. To put in more technical terms, the properties which are alleged to be mental properties by externalist theories are *epiphenomenal*. Everything in the world would happen in exactly the same way even if there were no such properties. Therefore, theoretical foundations of a mentalistic science of behavior cannot be externalistic, for possessing such properties do not contribute to the causal flow and nothing in the world can be causally explained by them. The externalist research program, which has been pursuing the second aim of cognitive science to help the field in developing

[2] Here, "world" should be understood in its technical sense, meaning "reality" or "actuality".

the theoretical foundations in order to properly pursue the first aim, is a dead-end.

This claim is not new. However, the discussion of the problem mostly came to a halt in the early nineties. Most of the literature in the history of the discussion was produced as philosophers spontaneously replied to each other and to the developing problems of their time, the productive period of the discussion coming to an unresolved halt in the past decade without any work that is looking back at the issue from a distance to bring the pieces together. No book-length or thesis-length work exists that discusses the intricacies of the whole problem.[3] Also, there are many versions of externalism and many nuances about causation, and an exhaustive treatment has to do justice to these matters. This thesis will hopefully be an example of such a work as a step to bring the problem closer to resolution, in favor of the view that externalism is not compatible with mental causation.

The problem is important both for scientific practice and philosophy of mind. Adopting a faulty theoretical framework might have negative practical consequences for a scientific paradigm, for its predictive and explanatory power, and its reputation in the intellectual world, and finally for its ultimate fate. According to many, behaviorism and psychoanalysis are such examples from the recent history. What troubles externalism may bring to mental and behavioral sciences, if it is really incompatible with their basic tenets, is an interesting question that I do not want to deal in

[3] An exception is Newman 2002. However, Newman devotes a lot of space to vindicate internalism against externalism, which will not be done here, and he doesn't consider many intricacies of the problem, such as the objections raised to defend the causal status of e-properties. Also, Newman defines internalism in a way that I would avoid.

depth here now. I just want to point to the fact that one prominent externalist, Hilary Putnam, declared in his 1997 that if externalism is true, then cognitive science is at best a utopian enterprise.

Although the incompatibilist threat is regarded as an important challenge for externalism by many, including many externalists, my personal experiences with people working in the field reveal that the problem is hardly even recognized by many people, including many externalists. Also, very few practicing scientists are aware of the views that they implicitly adopt about the mind and their implications for their practice. Hopefully, the current work will be helpful in drawing attention to the problem, and hopefully it will be one that can be penetrated easily by non-philosophers.

The first chapter begins by defining our questions and methodology. Our question will be defined as “do e-properties bestow any causal powers on the entities that have them?”, and we will also introduce the somewhat minor question of “are e-properties causally relevant to the powers of the entities that have them, even if they are not causally efficacious?” The use of the philosophical method will also be justified. In section 1.2, I will begin outlining the theoretical and historical background of the problem, giving what I take to be the correct definitions of internalism and externalism, and introducing conceptual distinctions such as constitutive and causal dependence. Section 1.3 will introduce the reader into the history of externalism. The section can also be seen as an attempt to give a historical analysis of how developed externalist theories in philosophy of mind were developed from the so-called “linguistic externalism”. 1.4 includes another quick look at our list of e-properties in the light of the preceding sections.

The second chapter will develop a causal-explanatory case against externalism. After a quick look at the earlier worries about the causal inefficacy of e-properties in the literature of the past few decades, I will carefully build up my own case in section 2.2, where what I call “the method of comparison” will be more clearly defined and applied. Section 2.3 will include a clear map of possibilities that shows where we can go from here and how externalism can have a chance to defend itself. In section 2.4 I will consider the objections to the theory developed, taking a look at attempts in the literature which tried to show how externalism can be compatible with mental causation. These attempts, among others, include the so-called “overdetermination response”, “the action-theoretical strategy”, Dretske’s “theory of structuring causes” (Dretske 1988), Burge and Baker’s “explanatory strategy” (Burge 1993, Baker 1993), Jackson and Pettit’s “program explanation” theory, (Jackson & Pettit 1990) and an argument by Fodor as articulated in his 1994.

The third chapter is a summary of the thesis, written in a somewhat less technical language.

CHAPTER I

THEORETICAL PRELIMINARIES

1.1 Defining the Question and the Methodology

The question we are concerned in this thesis is the following:

(Q1) Do e-properties bestow any causal powers on the entities that have them?

Before making an attempt to answer the question, I believe we should ask some questions about the question itself. What kind of a question is this, and how is it to be answered? What do we mean by "a property bestowing a causal power on something"? Is the question to be answered by empirical or non-empirical methods?

First, we will try to understand what we mean here when we say that a property bestows causal powers on an object. The nature of causation is a question that hasn't been settled since the earliest days of philosophy and probably it won't be settled anytime soon either. However, even though there is no agreement on what causation *is*, I believe that we have more-or-less agreed upon methods to *look for* causes, which will be enough for the present work. One common-sensical idea that will be in the background of this thesis is that *if something is a cause, it makes a difference in the world.* Another will be an idea which has been emerging from the interventionist theory of causation in the last twenty years, which emphasizes the fact that *if X is causally relevant for Y, then one can in principle intervene with the occurrence of X to intervene with the occurrence of Y.* Interventionist theories of causation are believed to go along well with the practice of scientific experimentation. We intervene with something to see if it makes a difference.[4]

The difference-making understanding and the interventionist understanding of causation might be expressing the same facts in different ways. For our purposes this is not important, as long as they help us clarify our ideas about finding out causes. To repeat, for our current purposes we are not interested in the ultimate nature of causation. All of the above are commitments about how to find out causes, and I don't mean to imply that the nature of causation is exhausted by the phenomenon of difference making.

A property "bestowing a power on an entity" does not mean that the property has a power which is then passed on to the entity.[5] Such an expression stands to emphasize the fact that it is *an object's or event's having a property at a time* that is important for causation, that *objects or*

[4] For interventionist theories of causation, see Menzies and Price (1993) and Woodward (2003).

[5] This point has been made by Walter (2009).

events cause things in virtue of the properties they have.[6] We generally talk about an object, an event, a process, or a state of affairs as being a cause. However, this loose talk has the potential to mislead us about causation. When we say that X is a cause of Y, and X is an M-type state or object by having M-type properties which we mention to refer to them, this doesn't mean that X causes Y in virtue of its M-type properties. Consider: I can say "My father caused the window to break", but this does not mean that his property of "being my father", the property that I use to refer to him, is causally efficacious in the breaking of the window.[7]

I will also assume that the terms "power", "disposition" and "ability" mean the same thing. As the word "power" has a somewhat positive vibe to it, I will prefer using the other terms in certain cases for stylistic reasons.

So far so good for causation. If we have agreed on what we mean by an object's bestowing a causal power on an entity, the next task is to find out how to test the causal efficacy of e-properties. When we want to test the causal efficacy of a property, what we do is experimentation. If we want to see whether a button's being on is the cause of the sound emitted by a machine, we turn it off to see whether the sound goes away. If we want to know whether having a certain type of gene is causally efficacious in having blue eyes, we interfere with the gene to see what happens. It is not

[6] I wish to note that I definitely do not want my arguments to rest on the metaphysics of properties. A property can be an attribute predicated of an object, some abstract universal that is tokened in an object, a way something is, a mode, a powerful quality, or the cement and stone of objects. I am not committed to any of these theories, and as I currently have no reason to believe otherwise, I believe that such issues do not matter for the forthcoming arguments and discussions. I also want to note that, if "having a property at time t" is an event, the property-view I adopt here will be compatible with the popular event-view of causation.

[7] See Dretske 1981, Grush 2001 and Humphreys 1999 for discussions related to this issue.

possible from the armchair to say whether that gene has a role in building blue eyes or not. But in our case, we are not asking a question about particular properties. We are asking a question about a whole class of properties, and we are asking whether these properties make any difference, anytime, anywhere.

The current inquiry is not in the empirical waters. It is a philosophical one. But how can we establish that a whole class of properties is causally irrelevant, through armchair methods? One thing we can do is to put our best knowledge of the world in good use and *come up with hypothetical situations, where, for all we know, things will happen in a certain way*, and where we can clearly see what properties are bestowers of powers or not, in a way that we can generalize. In this thesis, this will be done by comparing two hypothetical entities, which share all their properties except e-properties, and try to see if e-properties make a difference or not. Details regarding how this will be done will be provided in section 2.2.1, where we will put the method to work.

We should note that running a thought experiment may show us a fact about the world in two ways. As noted above, it can provide an opportunity for us to make use of our knowledge, and some of this knowledge might be empirically acquired. Also, thought experiments can show us *conceptual impossibilities*. It can be the case that not only e-properties do not make a difference to what happens in the world according to our best knowledge, but also their making a difference *doesn't make sense at all*, that we cannot coherently think of a case where something is an e-property but is also causally efficacious. Both possibilities will be considered in this thesis.

Above we have defined our first question. Later, we will see that we will need to deal with another question, which is:

(Q2) Are e-properties causally relevant to the powers of the entities that have them, even if they are not causally efficacious?

(Q2) is not as clear as (Q1), for “causal relevance”, a term of art we often encounter in the literature, is not as well understood as “causal efficacy”. Causal efficacy of a property is surely a species of causal relevance, but something can be causally relevant without being causally efficacious. But in what sense something is “relevant” for causation without being causally efficacious? The only senses I can make of “relevance” here are *epistemic* and *pragmatic* ones: Knowing those properties, even though they are not causally efficacious themselves, are epistemically valuable for they let us formulate law-like generalizations or predictions because the non-efficacious properties systematically co-vary with efficacious ones. The way to see whether e-properties are causally relevant would be by checking whether they systematically co-vary with causally efficacious properties.

I hope so far we were able to clearly define our questions and the ways of solving them, at least clear enough to be workable. Now let me explain how the case to be developed in this thesis is supposed to work as “a case against externalism.” Surely, we will try to build a case for incompatibilism, the idea that externalism is not compatible with the phenomenon of mental causation, if there is such a phenomenon. However, this does not falsify the externalist theory, since it is logically possible that externalism is true and there is no such phenomenon of mental causation. But if the epiphenomenality of mental states is absurd for us, then the current case can work as a *reductio* for externalism. A reductio does not falsify a theory, but it can make us think twice. If mental states being causally efficacious and the successful mentalistic causal explanation is something we are much more strongly committed to than we are to externalism, be it lay or scientific cases of mentalistic explanation, we have the grounds to doubt the validity of externalism. The current thesis presents a case against externalism not by falsifying it, but by raising doubts about it. This does not mean that we wouldn't doubt externalism if it weren't for causal reasons, for there is a host of other problems associated with externalism. We will not go deeply into these problems within the boundaries of this thesis, however.[8]

Let us mention that the problem of mental causation in the context of externalism is not the only causal problem about the mind. There are arguments to the effect that there is something problematic about the causal efficacy of mental properties in general, no matter whether externalism is true or false; (Malcolm 1968, Kim 1998) along with arguments to the effect that mental properties cannot causally explain behavior because having a mental property just means having certain behavioral dispositions; (Ryle 1949) and arguments to the effect that mental properties are provoked to justify or rationalize behavior, not causally explain them. (Bennett & Hacker 2003) We won't be dealing with these problems at all. The solution of these problems might or might not have a bearing on our current problem, and vice-versa. I simply assume

[8] For criticisms of externalism on non-causal grounds, see Loar 1988, Boghossian 1998, Segal 2000, Chomsky 2000, Farkas 2008; and also the discussions in collected editions by Pessin and Goldberg 1996 and Schantz 2004.)

that they do not, at least not to a degree that will affect the conclusion of this thesis.

Before passing, let me give some examples of mentalistic causal explanation to make the term more lucid for the readers. We explain people's behavior by relying on their beliefs and desires, such as when we say that Hazar stayed in the bar longer than he normally does because he wanted to hook up with a girl and he believed that the girl was interested in him. The effect that is explained may extend further than the immediate behavior; for instance, we can say that it was Hazar's beliefs and desires that caused the eventual pregnancy of the girl. It is not only willful behavior or actions that we explain through mental properties. It is sensible to say that someone was sweating because she felt fear, or she woke up screaming because she had a nightmare. Mentalistic causal explanation can also be encountered in cases like the explanation of placebo effect and psychosomatic effects, which are far from being actions, or even instances of behavior. Scientific explanations may also rely on sub-personal or sub-conscious mental states and mental structures, such as the sub-conscious mental states attributed by psychoanalysis or early stages of information processing attributed by cognitive science. Finally, in everyday life and science, we also talk about causal relations between mental states, such as when a thought triggers a memory. The latter examples also make us doubt the philosophical theories which tell us that mental properties explain only actions, or the theories which claim that mental properties are mentioned to rationalize or justify one's behavior.

1.2 Historical and Theoretical Background of the Problem

In this section, we will briefly go through the history of externalism. This will serve two purposes: One is to justify my claim that externalist theories do propose as the basis of the mind all and only what I have called e-properties. The other is to introduce the non-familiar reader into the internalism-externalism debate for a fuller appreciation of what is at stake. We will also take a look at how the debate relates to the problem of mentalistic causal explanation.

Before going on to define internalism and externalism, we should first note what it is a disagreement about. The disagreement between internalism and externalism is mostly about contentful mental states. A contentful state of mind is a state that we express in the form “Person P is X’ing [about/of/that] Y”, where X is filled by a mental predicate, such as “think”, “desire”, “imagine”, which are also called mental attitudes, and Y is filled by “anything you can think of”, such as “table”, “chair”, “Ahmet”, “Atlantis” or “I will be OK.” The Y is the “content” of one’s mental state, what is “in one’s mind”.[9] According to some theorists, there are also states of mind, which are defined not by their object or content but by their quality or character. Feelings of pain and moods are given as paradigmatic examples, and according to some all perception-like states have such qualities. But fully externalistic theories claim that all such states are actually intentional states since that seems to be the only way to “externalize” them, so it is

[9] When I use the word “content”, I am following the common usage in the literature, without being committed to the idea of linguistic or other types of “content” being the same kind of thing as “mental content”. I will also assume that the technical phrase “intentional state” is interchangeable with “contentful state”, and “intentional object” with “mental content”. However, I will try to abstain from using the phrase “mental representation” for I take it to be a more theoretically loaded and misleading term, although that terminology is widely used in philosophy of cognitive science and even used as default in many circles.

safe to say that the internalism-externalism debate is a debate about intentional states.

1.2.1 Internalism and Externalism Defined

Internalism and externalism is a disagreement about the foundations of mentality, that is, it is a disagreement about what kind of properties one needs to have in order to have a certain type of mental state. Internalism or externalism, by themselves, are silent regarding exactly what kind of properties mental properties are in this world; they make a more general claim about the nature of mental states. One can be an internalist or an externalist while being silent about many other things regarding the nature of mentality.

Although the names of the two theoretical positions seem to give us clues, defining internalism and externalism is not easy, and many popular definitions might be misleading. A way to distinguish internalism from externalism would be to say that internalism claims that mental properties are intrinsic properties, while externalism claims that they are relational ones, that is, one needs to stand in a relation to something to have a particular kind of mental state. To use the jargon of logic, one could say that internalism says that mental predicates are one-place predicates, while externalism says that they are two-place predicates. Even though this definition is not as misleading as the ones that we will encounter soon, it is far from being unproblematic, for it seems possible to have a two-place predicate holding while one of the two entities does not exist. Dispositional properties are examples. A sugar cube is disposed to dissolve in water, but if all water in the world ceased to exist, it still sounds sensible to say that the sugar cube is disposed to dissolve in water. It sounds sensible to say that a building has the power to survive a possible hurricane, while the "possible hurricane" does not or maybe even will never exist. This point is important because when we take a look at what externalist theories have been claiming, we can see that they have been claiming that one's having mental states depends on the *existence* of *concrete* things outside of her, now or in the past. If mental properties turn out to be relations to "possible" things, or abstract objects such as propositions or universals, it wouldn't be good news for externalism.[10]

According to one definition of internalism, a definition based on a popular early paper by Putnam (1975), the internalist claims that the properties relevant to mentality are exclusively found in the properties of the entities "inside the skin", while the externalist denies that. However, as Farkas argues, this cannot be right, because those mental states the content of which deals with things internal to one are also subject to a disagreement between the internalist and the externalist, even though all the relevant properties are literally inside the skin. One might make an attempt to transform the "inside the skin" criterion to an "inside the brain" one, but this still doesn't go into the heart of the disagreement between the internalist and the externalist, since it is possible to be an internalist while being a substance dualist who believes that there are non-physical parts of a person, other than the brain, that is relevant for mentality. (Farkas 2003, 2008) Plus, when we think about thoughts about one's own brain, the "inside the brain" criterion faces the same problems faced by the "inside

[10] Gertler (2007) also makes this point.

the skin" criterion. Moreover, the ideas behind internalism and externalism can be traced back to times when we didn't know that the brain is a crucial organ for mentality.

To avoid the above mentioned problems, maybe we should leave subjects or organisms aside, for they can think about their own parts, and their thoughts about their own parts are still a matter of debate between the internalist and the externalist. Maybe we should focus on not the whole organisms or subjects, but their minds (or individual mental vehicles), and what is inside or outside them. It is both possible to hear philosophers talking about subjects having mental properties, or equally, their mental states or their brain states having mental properties. These two ways of talking are compatible, as far as subjects have mental properties in virtue of having states, body parts or "soul parts" with mental properties. For example, "*having* a thought about tigers" might be a mental property of a subject, while "*being* about tigers" might be a mental property of the thought. The thought might be a state or a part of the subject or the organism, material or immaterial. This way of looking at things might help us define internalism and externalism in a better way without getting into "inside the skin" talk. If thoughts, desires and concepts are thought of as parts or states of the subjects, what externalism might be claiming at heart is that what makes these entities mental might be facts outside these entities, whatever these entities ultimately turn out to be, and no matter whether these external facts are inside the organisms or the subjects in a broad sense. Internalism is a denial of this position.[11]

[11] Two digressions before moving on: According to some philosophers, there is literally an object called the mind, most of those philosophers today also believing that that object is the brain. According to others, even though there is such a property as *having a mind*, such as having certain abilities, there is no such object as '*a' mind*, just like the fact that there is such a thing as having a power without there being objects called powers. The

Following what we have said above, I think the best way to define externalism will be thus: Externalism is the thesis that having a mental property constitutively depends on the existence of concrete things and their properties outside the mental vehicles of the subject, and certain relations between those vehicles and these things and properties outside those vehicles. Mental vehicles can also be called "core realizers", to follow Shoemaker's terminology. (Shoemaker 2007) Internalism is the denial of this position. As it should be clear by now, internalism defined this way is compatible with the possibility of mental properties being constituted of relations to *possible* or *non-existent* things, if such a thing makes sense.

This way of defining externalism also helps us to differentiate it from another theory that goes by the name of externalism. There is a theory on the market that is generally labeled "vehicle externalism" or "extended mind theory". (Clark and Chalmers 1997, Clark 2008) Although these theories are referred to as externalist theories, they differ from the theory of *content externalism* which is the target of the current work, so they

same goes for thoughts. It is possible that there are *events* of thinking, and such a thing as *having* a thought, but no such thing as *'a' thought*. Those philosophers who ascribe to the latter position will concede that there are objects, such as brains, that are crucial in endowing organisms with the property of having a mind and that differ from their other parts in this respect; and certain more specific brain events will be the subsumers of specific thoughts. To not to complicate the matters even more, I will kindly ask the readers who ascribe to the second position to let me call those distinctive parts of the organisms their "minds", and the specific states or events as "thoughts".

The second digression is about our talk of subjects' *having* minds. According to one philosophical tradition, which begins at least from Plato or Descartes and continues to most of today's neuroscientists who claim that "we are our brains", subjects are just their minds and nothing more. (See Bennett and Hacker 2003 for a historical analysis.) I don't know if that would simplify or complicate the matter of giving a definition of externalism and internalism if it were true. I just assume that for the current endeavor, its truth or falsity doesn't make a difference.

should be differentiated from it and put aside. Both positions can be loosely defined as the thesis that "mentality depends on things outside the subject", which puts the theories in the same basket, but this loose formulation would be very misleading.

Many content externalists, while holding that mentality depends on the relational properties of internal states, still claim that mental states themselves *are* internal states of organisms. They hold that mental processes are processes that are literally going on in the heads of organisms, even though what makes these processes mental are their relations to things outside the head. Many also hold that the mind is the brain, if the mind is an object at all. Vehicle externalists, on the other hand, hold that some mental processes take place outside the head, and the brain is not the only substrate of mentality. For instance, the position of Clark and Chalmers (1997) imply that when you are doing a calculation with pencil and paper, the pencil, paper, and the numbers written on the paper are literally parts of your mind, and the processes going on involving the paper, pencil and the paper are mental processes.

The debate between vehicle externalists and "intracranialists" is not a debate about what makes something mental or what fixes the intentional content of a mental state, but the *location* of the mind or mental processes. The "externalist" position I am targeting is the one regarding the constitutive properties of mental states, not their location. Therefore, vehicle externalism is not affected by anything I will say in the rest of the thesis.[12]

[12] I do not want to give the impression that I believe vehicle externalism is true or even plausible, I do not. What we need to agree on before seeing whether mental processes

Vehicle externalism was, unfortunately, labeled in a way similar to content externalism, for they were both based on the "outside the skin" idea. Vehicle externalism is externalistic in the sense that it points to the non-biological processes outside the organism. However, it is or should be committed to the thesis that these non-biological processes are parts of cognizing subjects. If a process going on in the pencil and the paper is a part of my mind, it is a part of *me*, and therefore such mental processes are *not* external to me. I take it that the central point about vehicle externalism has been that there can be mental processes outside the brain or the body. However, a content externalist cannot say that if my thoughts about tigers depend on the existence of tigers, tigers that walk this earth are literally parts of me. Not only because it sounds prima facie absurd, it would also make it hard to understand the claim of content externalism and how it differs from vehicle externalism. If tigers are parts of my mind, they are neither external to me, nor external to my mind. They are for sure external to my human body, but as we have seen, being external or internal to a body has never really been what is at stake between content externalism and internalism.[13]

Even though we have just done hard work to define internalism and externalism, it is possible that we may not need this definition for the

extend beyond the brain or not is a theory about what makes a state mental in the first place, and what makes some arguments for vehicle externalism look plausible on the first glance is just the fact that they lack such a theory and work with a very loose definition of the mental. See Adams and Aizawa (2008) for elaborations on this point and a rebuttal of vehicle externalism.

[13] The point that content externalism should define itself in a way that doesn't make it collapse into vehicle externalism has been urged by Gertler, who also presents us the hardness of this project. (Gertler 2007)

purposes of this thesis. It is also possible to define externalism by looking at what externalist theories have been particularly claiming, looking at the theoretical products of the history of the rather young tradition. Externalism is the thesis that having a mental state depends on possessing some or other e-property. E-properties are, to repeat:

Synchronic-Relational: Being in a world where there is X.

Historical-Causal: Being caused by X.

Historical-Use: Being a cause of X.

Normative: Having an etiological proper function.

Similarly, internalism can be defined in a negative way. Internalism is a denial of mental states' being dependent on e-properties. The above mentioned e-properties are surely relations to concrete entities that exist or have existed, so they fit into our definition. Anyway, I will prefer working on specific properties that have been suggested by externalists. This will also help us sidestep the issues about the metaphysics of intrinsicness and extrinsicness and the borders of a subject. All this, I hope, gives us some workable and substantial content to the terms "externalism" and "internalism".

1.2.2 Causal Dependence and Constitutive Dependence

I have said that internalism is a denial of mental states' being dependent on e-properties. Here, "dependence" should be understood as *constitutive dependence*, as opposed to *causal dependence*. The debate between internalism and externalism is *not* a debate about what mental states *causally* depend on, but what they *constitutively* depend on.

The case of statues is a good example to demonstrate the distinction. Statues seem to *both* causally and constitutively depend on human intentions. Statues causally depend on sculptors, which means that at least generally, you need a sculptor to produce a statue. Something that has the same shape as a bust of Lenin could *in principle* be formed by natural means without human intervention, but it is unlikely. But it also seems like the property of "being a statue" not only causally, but also constitutively depends on sculptors. A naturally formed rock, even if all of its relational properties are the same as a bust of Lenin, is not a bust of Lenin. So it seems like the relational property "being caused by human intentions" is a necessary constitutive property for something to be a statue. For the property of "being a bald head", the case is not the same as statues. There are many things that can turn a head into a bald head. However, what makes a head bald, in the constitutive sense, seems to be its intrinsic properties. A bald head, satisfying some intrinsic conditions such as "having no or little hair" is a bald head regardless of what caused the baldness, since these factors cause the very same thing, which is sufficient for a head to be called "bald". Although baldness causally depends on many things, it does not constitutively depend on any of them. It constitutively depends only on the number of hairs.

The internalist fully acknowledges the fact that in ordinary and biological circumstances, most mental states causally depend on external factors: Light is reflected off from surfaces, which causes some changes in my brain, the protein in the food I eat nourishes my neural cells, and environmental factors during evolution and development determine which internal states will stay and which will go. However, the internalist claims that what makes a state mental is to be found among the end products of these causal chains, and those end products suffice to make a state mental no matter what they are caused by or what exists outside them. The internalist thinks that having a mind is more like being bald than being a statue.

1.2.3 A Note on Dependence and Individuation

Sometimes it is possible to hear that what externalism claims is that some mental states are *individuated* in a relational way. (Sanford Goldberg, personal communication.) Individuation might be a misleading term, as at least one of the senses of the term can be understood as an epistemic activity, while dependence, constitution, and identity are ontological relations. If the claim of externalism was such that some mental states are individuated in a relational way, it would be a rather trivial thesis, since it is possible to individuate *everything* in a relational way. I can individuate my hands in a relational way, for instance: My left hand and my right hand. But nevertheless, the property of "being a hand" doesn't depend on such spatial relations. Similarly, to be an interesting thesis, externalism needs to say something stronger than that, something foundational about mental states, that one cannot have a mental state unless those relational properties are in place. After all, all internalists accept that mental states can be individuated in important relational ways. For instance, what makes my belief that there is full moon tonight either a piece of knowledge or misinformation depends on the existence (or the non-existence) of the full moon. However, all this is compatible with the two being the very same belief no matter the beliefs true or false, that one can individuate the very same mental state through their relations to non-mental facts, and that having the kind of belief does not depend on the existence of the moon or anything else other than my mind.[14] So, all through the thesis, I will assume that externalism is making this strong, foundational claim rather than the one about individuation.

1.3 A Short History of Externalism and How it Relates to the Problem of Mental Causation in Philosophy of Cognitive Science

Until recently, internalism was the dominant theory both about the mind and language in Western philosophy, even though the label "internalism" was not used, for there wasn't really any other position to contrast. A reason for internalism's popularity in philosophy of language was the simple observation that what determines the meaning of a word depended on the mind of the interpreter, a fact that is supported by the arbitrariness of linguistic meaning and the phenomenon of non-referring words. For mental states, many similarly held that they can stay exactly the same even though the world could dramatically change, and idea not only supported by Cartesian evil-demon style thought experiments but phenomena like false beliefs, hallucination, imagination and creativity, not to mention the effect of the popularity of subjectivist trends in philosophy of perception and psychophysics since the early days of modern science. Today in neuropsychological circles, it is also common to come across empirical arguments in favor of internalism.

[14] Beliefs can be individuated relationally through very trivial ways as well: My belief that there is wind outside is a belief that is occurring a hundred kilometers away from Bremen, but Tilman's belief that there is wind outside is a belief that is occurring a hundred and one kilometers away from Bremen. Although it is possible individuate these beliefs through their relations to Bremen, no one would say that their contents depend on their distance from Bremen when they occur.

Internalism has been strongly challenged lately, a challenge that mostly began in the seventies, though some of its roots might be traced earlier, maybe to the works of Wittgenstein (as seems to be suggested by Heil 2004) Externalism became immensely popular as a result of certain thought experiments which allegedly show that our intuitions support externalism, and these intuitions were later used to find solutions to a problem in theoretical cognitive science, the problem of naturalizing mentality.

1.3.1 Cases for Linguistic Externalism

The key factor behind the spread of externalism in the recent history of philosophy of mind is definitely the development of linguistic externalism in the seventies. Below we will take a look at those famous thought experiments in favor of linguistic externalism, and then try to see how people moved to mental externalism from these considerations about language. I assume that many readers are familiar with them and I apologize for making the reader go through them once again.

1.3.2 Kripke's Gödel

The first one of the aforementioned thought experiments is Kripke's "Gödel" thought experiment. (Kripke 1972) It was designed to dethrone the popular descriptivist views in philosophy of language, which stated that the meaning of a word is fixed by a description in the speaker's mind.

Kripke wants us to imagine a case where it turns out that it wasn't Gödel, but someone called Schmidt who found the proof of incompleteness. If, Kripke says, descriptivism were right and the meaning of the word "Gödel" were fixed by a description like "the person who discovered the proof of incompleteness" then it would imply that in such a case we had been talking about Schmidt all the time, for he was the one who discovered the proof. However, that's absurd, for our word "Gödel" refers to Gödel, not Schmidt, even if it turns out that he wasn't the one who came up with the proof.

The moral Kripke brought home from this thought experiment was that the description one has in mind does not fix the meaning of a word, at least for proper names. Instead, he proposed a *causal theory of reference*, where facts external to the speaker's mind fix the meaning of a word. For Kripke, these facts were facts about Gödel's being "baptized" by the name "Gödel" and a causal chain of linguistic use leading from that event of baptizing to the utterance of the word by individuals today.[15]

1.3.3 Putnam's Twin Earth

The more famous and influential of the early thought experiments in support of externalism and against descriptivism was definitely Putnam's Twin Earth. (Putnam 1975) I assume that many readers have heard the story dozens of times, but it is useful to retell it since we will later use the very same scenario while building a causal case against externalism.

[15] The descriptivist response to Gödel was *causal descriptivism*, the idea that the related descriptions also include causal facts. For causal descriptivism, see Jackson (1998), Kroon (1987) and Lewis (1984).

Putnam wants us to imagine a planet which is exactly the same as our planet, except for the fact that wherever we have H_2O on Earth, there is a different liquid on Twin Earth which he dubs "XYZ". These planets exist in the same universe, and the setting is a time before modern chemistry. The inhabitants of both planets are oblivious to the facts about the microchemistry of these liquids. They both use the word form "water", and when asked what water is, they naturally give the same description, such as "it is a potable, transparent liquid that falls from the sky and flows in the rivers."

Putnam's intuitions, which are shared by many contemporary philosophers of language, say that when people on Earth utter "water", it means H_2O, not XYZ. When people on Twin Earth utter "water", it means XYZ. If someone from Earth goes to Twin Earth and says "this is water" pointing to a pond of XYZ, he is saying something wrong.[16]

If these intuitions are correct, descriptivism seems to be false, for the only difference between these two planets is the chemical composition of watery stuff, which is a fact external to the mind of the speakers.

1.3.4 Burge's Arthritis

Putnam developed his case only for the so-called "natural kind terms". Burge, in his thought experiment where he tried to establish "social

[16] This part is a bit tricky, for some externalists say that when one goes to Twin Earth and keeps uttering "water", one is now talking about XYZ.

externalism" or what he called "anti-individualism", tried to generalize linguistic externalism for all kinds of terms, and tried to show that the meaning of the words one uses depends on facts about other people. (Burge 1979)

Burge invites us to imagine a speaker, Alf, who believes that he has arthritis in his thigh. Alf lives in a linguistic community where experts define "arthritis" as a disease in the joints. When he goes to the doctor, the doctor tells him that he cannot possibly have arthritis in his thigh, arthritis, by definition, is found only in the joints. On the other hand, we have Twin Alf, an intrinsic twin of Alf, and he also thinks that he has "arthritis" in his thigh, but he lives in a linguistic community where "arthritis" is defined as a disease that can be found also in the thigh. Note that the description related to the word form "arthritis" in the mind of Alf and Twin Alf is the same, they both include the fact that arthritis can be found in the thigh.

According to Burge, when Alf says that he has arthritis in his thigh, he says something wrong. Moreover, he is making a logical mistake. When Twin Alf says it, he is right. Therefore, the word form "arthritis" has two different truth conditions, and therefore two different meanings in the mouth of two speakers, and the only difference between these two speakers is their social community. Therefore, linguistic internalism, or "individualism" as Burge calls it, should be false.

1.3.5 A Route from Linguistic Externalism to Mental Externalism

There are two routes from these linguistic cases to externalism about mental states. One line, which was taken by Putnam and Burge themselves, goes like this: If people on Earth and Twin-Earth mean different things, then their utterances express *different concepts*. Therefore, they not only mean different things, but they also think about, or desire, different things with their "water" thoughts. This move from linguistic content to mental states might be logically correct or not, but nevertheless, it is now already common to run these thought experiments directly in terms of mental states rather than linguistic content. We will see the second route when it will become more relevant in the next section.[17]

These thought experiments didn't lead to an extreme externalism initially. Putnam himself noted that there is a kind of content that is shared by people on Earth and Twin Earth. This kind of content came to be called "narrow content", as contrasted to "broad content", and people who ascribe to there being two kinds of content came to be called "two-factor theorists." More extreme forms of externalism emerged as more

[17] Let me reflect on a little bit about how these thought experiments are meant to support externalism. They are, in no way, a vindication of externalism. The only thing that they can show by themselves is that the intuitions of most western philosophers are externalistic. For all we know, those intuitions can be false, and there can be another, non-externalistic way of making sense of mental states, which might be the true way of making sense of them. It is also possible that the thought experiments do not even unearth intuitions, but create confusions through the ways they are told. To vindicate externalism, the externalist should do something stronger; she should show us that there is no non-externalistic way of making sense of mental states. Unfortunately, the arguments have been sold too quickly to critically scrutinize the claims. One simple line of thought that can go against these ways of vindicating externalism, for instance, is to say that it is pretty well conceivable to imagine a person who has thoughts about water in a world where there is no water, just like people in our world have thoughts about unicorns. (Boghossian 1997, Segal 2000) So it is also possible that in our world, even though there is water, it is possible that having a water-thought doesn't constitutively depend on one's relations to water. The aim of this thesis is not to argue against externalism on these grounds, the reason I am mentioning these problems is just to make the unaccustomed reader aware of the fact that the popularity of externalism depends at best on some historical factors, not its being a well-worked and well-scrutinized theory.

sophisticated naturalistic theories were developed to solve problems regarding the metaphysics of mind and theoretical cognitive science.[18]

1.3.6 Externalism and the Naturalization of Mental Content

The construction of fully developed positive theories about the externalistic foundations of mind from the above negative theories against internalism began as some problems of theoretical cognitive science needed an urgent solution.

In the beginning, I said that contemporary cognitive science is marked by, at least, two aims: The first one is explaining behavior through mental states and properties possessed by living things, and the second is figuring out the place of such states and properties in the natural order. The second aim is pursued in philosophy as an attempt of giving a *naturalistic constitutive analysis of mental states*. What I mean by a "constitutive analysis of mental states" is a theory that defines the minimal sufficient properties had by a state at a given time for that state to be a mental state, properties which themselves do not include mental properties.[19]

[18] Even though the two-factor theory is rather common and is taken to be a sensible position to hold, it is not wholly clear what is meant by there being two kinds of content. After all, when I think about water, I have one thing in mind: water. The two-factor theory is better interpreted as saying that mental states are determined by narrow and broad factors, which might again turn out to be problematic when we begin considering whether one can have one without the other. Anyway, we should not strongly expect to find answers to these questions in the literature, since the two-factor theory developed spontaneously and unhesitatingly as a result of philosophical pressure, without proper initial scrutiny.

[19] Surely, it is possible to analyze a mental state in terms of *other* mental states; however, this will not be an analysis of mentality in general.

What I mean by “naturalism”, in the context of philosophy of mind, is the thesis that mental phenomena are constituted solely by phenomena that are known to or not contradictory to natural sciences, the fundamental natural entities not being mental.[20] Most self-proclaimed analyses we encounter in the literature go in the form: “Mental states are X”, where X is filled out by properties that are themselves alleged to be non-mental, and which are found among the phenomena known to natural sciences.[21] The question why anyone would want to be a naturalist, although being an interesting question, shall not trouble us now.

The aim of giving a naturalistic constitutive analysis of mental states in cognitive science circles becomes especially apparent when the project is welcomed by many as a hope of a “full naturalization” of the mind, and when the new notions and new theories that are introduced are seen as useful tools for this naturalization, tools for providing answers to

[20] Naturalism in this sense should be distinguished from what Chomsky has called “methodological naturalism”. (Chomsky 2000) One can choose to study the mind with the methods of natural sciences, seeing the mind just like any other object in the natural world, without expecting mental predicates to be analyzed into non-mental predicates, or without expecting the mind sciences to be harmonious with the current non-mental sciences. Also, the rejection of these forms of naturalism should be called “non-naturalism”, rather than “supernaturalism”. One can hold that mental phenomena are neither constituted by non-mental phenomena, nor can they be studied as a natural object, but this doesn’t lead one to claim that there is anything supernatural (e.g. intervention of a divine agent) involved in things mental. We might also need to differentiate “causal naturalism” from “constitutive naturalism”. Some versions of creationism (e.g. Plantinga 1993) might be constitutively naturalistic without being causally naturalistic, if they hold that mental properties are constituted by natural ones even though they cannot be caused by natural environments without some supernatural intervention.

[21] This last criterion deserves some attention, though. The constitutive properties offered by naturalistic theories of analysis are very rarely culled from the findings of natural sciences. Rather, the theories rely on general metaphysical categories like causation and co-variation, which are somehow labeled “natural”. So it is better to view this project as finding a *topic-neutral* analysis of mentality, that is, an analysis through properties that are common both to the mental and the non-mental discourse.

challenges posed by classical or contemporary philosophers. These theories include “computational-representational theory of mind”, “information processing account of adaptive behavior”, and “symbol manipulation theory of cognition.”

The mental states of modern cognitive science in their technical dress-up still look like everyday intentional states: Information and symbols are “about something”, computations and representations are computations and representations “of something”. However, these states continue to be theoretically problematic. Currently, there is no agreed upon answer regarding how to objectively determine the right hand side of these “ofs” and “abouts” for a given system, or which systems have these kinds of states, or even whether such states exist or not. There is no agreement about the minimal sufficient natural conditions for having these states.[22]

Most responses given to the problem of providing such an analysis of intentionality has been dominated by externalist theories. One important motivation for externalism was naturalism, because many people have been thinking that the innards of organisms, particularly their non-relationally understood neural properties, provide no basis for such a constitutive analysis or an objective way to determine the content of an organism’s mind. Wittgenstein’s arguments to the effect that it is impossible to determine what one means by one’s words or what one has in one’s mind by looking at the internal states of the individual had

[22] The reader should be mindful about the possibility that by trying to give an account of mental states in terms of other intentional states, informational theories in cognitive science might have just postponed the problem in an unnecessary way. The worse possibility is that these theories confuse us in thinking about the mind, rather than aiding us in understanding it.

probably prepared the ground for such an idea, supported by the more recent arguments by influential philosophers like Stalnaker to the effect that it is not possible to make sense of internalistic content. (Stalnaker 1989; Burge 1982 includes similar remarks.) The other important motivation was surely the popularity of linguistic externalism, spread by the above mentioned thought experiments by Putnam, Burge and Kripke.

This led to a research project being culminated for the last forty years or so, characterized by an attempt of trying to find a constitutive analysis of mental properties relying on natural relations between the internal states of organisms and their environments. Following Kriegel and Horgan (forthcoming), I will call this project *The Naturalist-Externalist Research Program*. (For representative examples of The Naturalist-Externalist Research Program, see Dretske 1981, 1988, 1995-a; Millikan 1983, 1994; Papineau 1993, Fodor 1987, 1994, Tye 1995)

1.3.7 The Route from the Representational Theory of Mind to Externalism via Linguistic Externalism

The use of semantic notions such as *representation*, *meaning* and *symbols,* and hypotheses like *the language of thought* (Fodor 1975) in cognitive science made philosophers put the mind in the same ontological bag with public language, which was itself the result of a huge impact of the idea that *mental processes are computational states,* or *the manipulation of entities with semantic properties*.[23] The other factor that made people put

[23] This idea is generally attributed to Alan Turing. It is true that Turing provided an abstract proof of the possibility that if we can formalize mental processes in a computational way, then we can, in principle, build machines that can behave like a human being. However, it

the two phenomena together was the "observation" that both words and thoughts are *about* something, and they both seem to have representational properties. Accordingly, many thought that whatever applies to language applies to the mind. Linguistic externalism was already in the air when these ideas became popular. Hence, the second route from linguistic externalism to mental externalism.

Let us remember that linguistic externalism says that a word's having a meaning is a relation between a word, a user, and something else in the world (even though the role of the user is less clear than the traditional accounts.)[24] Mental externalism, built upon the above mentioned ideas, states that what makes a state in the brain a mental one is a relation between that state and something outside that state for it to be a symbol.[25] It is an indisputable fact that the meanings of words are not determined by the intrinsic properties of the words, such as the vibrations in the vocal chords or the chemical make-up of the ink. Therefore, externalists argued, the meaning of mental representations, mental

is unclear whether Turing has said anything to the effect that mental processes *are* computational processes. His 1950 is clearly dominated by a behavioristic vision, and according to behaviorism, the criteria of mentality is behavior, not the possession of symbolic states. It would be a more historically correct analysis to see Turing as offering his ideas about computation as a way to build thinking artifacts, not as a criterion for mentality. However, this minor historical correction may have scandalous effects on cognitive science, for if it is true, a large portion of the theoretical background of cognitive science rests on a misreading of Turing. When we realize the possibility that computational formalization being a way to build machines that think or simulate thinking doesn't imply that mental processes *are* computational processes, it is unclear what motivation is there to believe that mentality is the possessing or manipulation of semantic entities.

[24] Burge-style social externalism also includes other users in the picture.

[25] In the early days of externalism, "the user" was not in the picture, but as we shall see, some problems forced the theorists to introduce something like a user.

representations being brain states, cannot be fixed by their intrinsic properties either.[26]

1.3.8 The Causal-Informational Theory of Mental Content

Among the first externalistic attempts to naturalize mental states has been the *causal-informational theory* of mental content, also known as "indicator semantics." Basic sketches of a causal theory have existed in the early arguments of Putnam and Kripke. Although the work of Field (1978) is also an example of an early attempt, it was Fred Dretske who popularized the theory in a series of influential works. (Dretske 1981, 1988, 1995-a) Informational theory of mind is especially important in the current context, as most cognitive scientists and neuroscientists seem to adopt the term "information" in their dealings with the mental.

The idea behind the causal-informational theory is rather simple: Mental states are states that carry information; something's carrying information about another thing is a matter of that thing's being caused by that other thing.[27]

[26] Even if it is true that mental phenomena are a species of representational-symbolic phenomena, it still doesn't follow so easily that their meaning cannot be determined by intrinsic factors, for brain states have important intrinsic properties that ink doesn't. By having certain brain states, one is intrinsically disposed to behave in certain ways and possess some capabilities, while we cannot say the same thing for the ink and the paper. Moreover, even if it is true that brain states, as vehicles of representation, need to stand in a relation to something to be representations, it doesn't follow that these things should be things in the environment. It is pretty well possible that the relations are relations between other brain states or the whole organism.

[27] Why anyone would have any motivations to think that mental states are information carrying states is another story, a story that is, at least for the current author, hard to uncover.

The earliest causal-informational accounts of mental content faced certain important problems. One of them was *the problem of indeterminacy*. Causation is a transitive phenomenon, so the brain state we are trying to determine the content of is caused by very many things. Another problem of indeterminacy is one that arises from co-instantiated properties so that the brain state always causally co-varies with both, such as "being a rabbit" and "being a collection of undetached rabbit parts".[28]

The other and more important problem was *the problem of misrepresentation*: A crude causal theory doesn't seem to leave a place for misrepresentation, since there is no such thing as "causal misinformation". One can see a dot on the wall as a fly, although the mental state, *seeing X as Y*, is caused by the dot. According to the crude causal theory, the brain state represents whatever causes it, leaving no place for illusions, hallucinations and false beliefs.[29] It accounts for "representing", but not "representing *as*".

Faced with this problem, the theorists tried to refine the theory, for example, by saying that a state is a mental state if it is not caused by a certain thing, but caused by it in a *reliable* way, or if it is triggered by something in *optimal conditions*. However, it seemed to many to be pretty

[28] The example is from Quine, who was aware of this problem, in a slightly different context. (Quine 1960) See Fodor (1994) where he transforms the Quinian problem of reference to a problem of mental content.

[29] It is generally cases of error that are discussed in the literature, however, it seems clear that the theory also has troubles with many non-erroneous cases such as imagination, hypothetical thought and creativity.

obvious that this optimality and reliability cannot be spelt out in a non-circular way, without recourse to further mental states.[30]

Another obvious problem for informational semantics has been the fact that causation is everywhere. There are so many things in this world that causally co-vary with each other. This was another factor for informational semantics to introduce other factors, if we don't want to accept the absurd conclusion that mental states are everywhere, if we don't want to be "pansemanticists".[31]

These problems led to the introduction of other factors, which would play the role of the "user" who determines the meaning of a mental symbol.

1.3.9 Teleosemantics and Use-Theories of Mental Content

It was evident that the causal-informational factors were not enough to fix the content of a mental state, and this is where other considerations were brought in. It looked like there was the need of a natural analogue of *norms* that could fix the content of linguistics units, something natural that should tell us not only what some brain states indicate, but what they *should* indicate.

[30] See Loewer 1987 for discussion. Also see also Fodor's asymmetric dependence theory, which tries to find a solution to this problem, and which for some reason, didn't become popular in the philosophical community. (Fodor 1987, 1994)

[31] Thanks to Erdinç Sayan for the term.

Teleosemantics is one such attempt to find such natural norms. According to teleosemantics, certain events in the biological history of organisms endow their parts with *proper functions*. The proper function of a body part is not what it does, but what it *should* do. These events are events that contribute to the survival and/or propagation of the organisms or the part that is performing the function. According to teleosemantics, brain states are naturally selected in virtue of their indicating things in the environment, such as there being a state of neural firing when there is a certain property in the environment. [32] Teleosemantics views misrepresentation as a case of *malfunction*, an organ attempting to indicate something at the wrong time.[33] All this, of course, depends on the assumption that it makes sense to speak of natural norms, that it makes sense to speak of what the parts of an organism should do without there being any minded entities that judge things to be good or bad.

Teleosemantics is a subset of *use-theories of mental content*, theories which say what a state itself causes is among the factors that determine the content of a state, and not (or not only) what causes the state. Not all use-theories are interested in biological history of organisms, but as there are very many things caused by bodily states of organisms, other use-theories also need to introduce some normative criteria. This is possible, for instance, by focusing on the behavior that fulfills the *needs* of the organisms. These theories overall, can also be called "consumer semantics", for they rely on the mechanisms that "consume", or make use

[32] Ruth Millikan is generally regarded as the mother of teleosemantics, and Dretske adopted the idea to complement his causal-informational theories. (Millikan 1987, 1993; Dretske 1995)

[33] It is unclear what teleosemantics has to do with cases like imagination or thought, given that these states are not states of misrepresentation so cannot be thought of as a malfunction, but they work in cases where the thing represented is not around.

of brain states.[34] Brain states are symbols, and the organism is the "user" that determines what they mean.[35]

Teleosemantics offers a solution to the problem of misrepresentation: Once the content of a brain state is normatively fixed, then it has that content whatever it is caused by.[36] It offers a similar solution to the problem about the transitivity of causation. Nevertheless, it has other problems which are yet to be solved, which we will not go into here.[37] The aim of the current section is illustrative, and it is beyond the current scope to criticize these theories on non-causal/explanatory grounds.[38]

1.3.10 Other Motivations for Externalism

[34] The term "consumer semantics" was, again, coined by Millikan (1984).

[35] For more on use-theories of mental content, see Bermudez 2003.

[36] It is not very clear how causal co-variation plays a role after the introduction of factors which are related to use. Dretske still focuses on the importance of co-variation, while Millikan seems to think that it is not a necessary factor.

[37] See the contributions to Macdonald and Papineau (2006), especially the introduction by the editors.

[38] Here, I just want to point to one general problematic aspect of the whole program, which will also help the reader to get a simpler picture of what's happening. I hope the general framework of the Naturalist-Externalist Research Program is by now clear: The program rests on the assumption that mental content is a species of linguistic content, and its strategy is to find natural analogues of what we find in languages: Signs, together with the intentions and norms behind their use. However, intentions and norms are mental entities, but the program's aim is to analyze mental properties into non-mental properties. It looks like the case that the program is trying to achieve an impossible goal, and there might be something wrong with the assumption that mental content is just like linguistic content.

There have been other arguments and motivations for externalism, which I will just deal with in a few sentences here because they were not as pervasive as the others. Some of them were epistemic, fueled by worries that only externalism can guarantee that we know something about the world.[39] Such worries are also apparent in the defense of *disjunctivism* about perception, the rejection of common-factor theories of perception which state that there is something mentally common between veridical perceptual states and states like hallucination, illusion and imagination. Most disjunctivist theories state that when having a veridical perceptual experience, the object which is the cause of the experience is something that determines what kind of mental state the state is.[40]

There have been certain arguments, most famously given by Evans and McDowell, to the effect that for thoughts involving non-descriptive proper names and demonstratives like "this" and "that", there cannot be a thought unless there is an entity that corresponds to the content of the thought. (Evans 1982, McDowell 1977) However, it is safe to say that these ideas are generally taken to be rather extreme by many and are not very popular.

1.4 E-Properties Redux

[39] Bilgrami 1992 and McDowell 1994 constitute good examples for such worries.

[40] Common-factor theories are not necessarily internalistic, but disjunctivism seems to be necessarily externalistic. For disjunctivism, see the collected edition by Byrne and Logue (2009), particularly the introduction of the editors.

As we are done with the overview of externalism, let us again return to our list of e-properties, the properties which are claimed by externalist theories to be the constituents of mental properties, or what mental properties are identified with:

(Synchronic-Relational) Being in a world where there is X.

(Historical-Causal) Being caused by X.

(Historical-Use) Being a cause of X.

(Normative) Having an etiological proper function.[41]

The synchronic-relational properties are suggested by, according to one reading, Burge and (according to a synchronic reading) Putnam style externalism, disjunctivist theories of perception, and Evans-McDowell style externalism for demonstratives and proper names. Historical-causal properties are suggested by Dretskean informational semantics, Kripkean causal theory of reference and (according to a causal reading) Putnam style externalism. Historical-use properties and normative properties are suggested by Millikanian and Dretskean teleosemantics and use-theories of content. For our purposes, normative properties can be grouped under historical properties, for according to the externalist theories we have seen, they are identified with historical-use properties.

[41] This also seems to be a good occasion to make a terminological distinction between *synchronic* externalism and *diachronic* externalism. A diachronic externalism need not be committed to the idea that relations I have with things that exist right now are relevant for my mental states, while the synchronic externalist is committed to it.

I hope I have now justified that my selection of e-properties are not ungrounded and that the selection really reflects the theories I will be arguing against.

Now, we can turn to the earlier worries about the causal inefficacy of e-properties.

CHAPTER 2

THE CASE FOR THE CAUSAL INEFFICACY OF E-PROPERTIES

2.1 The Earlier Worries about the Causal Inefficacy of E-Properties

We have said that if it turns out that e-properties do not bestow any powers on the entities that have them, then the legitimacy of a cognitive science or psychology backed by an externalistic philosophy is to be doubted. Right after the spread of externalism, many philosophers started to have worries about the causal efficacy of e-properties, important proponents of externalism notwithstanding.

Before documenting these worries, I apologize to the reader for making her go through one last necessary clarification about the problem.

The question of e-properties being causally relevant or not is *not* a question about the *properties of the environment* being causally relevant or not. It is obvious (*pace* idealism) that things in our environment produce changes in our body, and given the transitivity of causation, they are among the causes of our behavior. The same thing can be said about the events that happened in the past. What is at stake is not whether the properties of the objects and events around us bestow causal powers on *them*, but the causal efficacy of *our* e-properties, which is supposed to bestow causal powers on *us* at any given time.

We have noted before that there have been many two-factor theories of mental content, dividing mental content into *narrow* and *broad* content. After the spread of the popularity of Twin Earth style externalism, many noted that it is the narrow, not broad content of one's mind that explains one's behavior. If I believe that water is potable but you don't believe that it is potable, everything else being equal, I will drink water and you won't. Both of our beliefs are about the same things, and according to externalism, they are about the same thing regardless of how we conceive of water. According to two-factor theories, what we think about is a broad fact, and how we conceive of it is a narrow fact, and, at the first glance, it seems that it is the narrow state of affairs that determine how we will behave. Moreover, there are the so called "Frege cases", where two thoughts pick out the same entity in the world, such as thoughts about the Evening Star and the Morning Star. Reserving a place for narrow content and treating these co-referring thoughts as two different kinds of thoughts help us a lot about predicting the behavior and mental transitions of the subjects. (See e.g. Fodor 1987, Rey 1997, Loar 1988, Segal 2000)

Another worry about externalism, especially of the historical sorts, was fueled by an imaginary creature called "Swampman". Swampman is a microphysical duplicate of a person who lacks all the relevant historical properties that person has. As Swampman has all the physical properties its twin has, we would expect him to behave in the same way in the same circumstances. However, historical externalism denies mentality for Swampman. But, the argument goes, as he behaves in the same way, mentalistic explanation should cover both Swampman and his twin. For if there is something that explains the behavior of Swampman, it will explain mine too, and if that explanans is not mental, mental properties will not figure in the explanation. We will deal with Swampman in more detail in section 2.2.1.

Some worries about the causal inefficacy of e-properties in matters mental were based on general considerations about the causal inefficacy of these properties in general. For instance, the historically important discussions in Dretske (1988) and Fodor (1987) try to see whether relational properties can be causally relevant in general, given the apparent fact that causation works by a transitive modification of objects' non-relational properties and that there are always some non-relational properties running the causal business.

One of the earliest expositions of such a problem can be found in Dretske (1988, p. 79) who provides the useful analogy of the opera singer: When an opera singer sings "shatter!" in a very high note, the glass shatters. What makes the glass shatter is not a relational semantic property of the word, but its non-relational physical properties. Dretske leaves us with the following problem: In other cases where we assume that semantic properties of words are causally relevant, such as when people understand words, how can they be causally relevant given that there are always some non-relational physical properties of the sounds doing some causal work, such as affecting the ears which then affect the brain. And how are we going to solve the problem for the semantic properties of mental representations?

We will see in section 2.4.6 how Dretske himself tried to solve the problem. However, not everyone thought that it is solvable. Stich, for instance, concluded that if e-properties are causally irrelevant and if the content of a mental state can only depend on these properties, then it is possible that these mental states don't exist at all, or at least, they have no use in scientific explanation. (Stich 1983) That is, some chose to get rid of mental content altogether instead of going for a two-factor route.

Of course, the route other than the two-factor route is a pure internalism. (See e.g. Chomsky 2000, O'Brien and Opie 2004, Segal 2000, Grush 2001 among those who also rely on causal considerations in defense of a pure internalism.) However, that route was rarely taken.

However, some didn't give up, and continued to argue that despite the initial appearance, e-properties *are* causally relevant after all. We will take a look at those responses to the problem in section 2.4, after I present my own way of establishing the causal irrelevance of e-properties.

2.2 Causality Check for E-Properties

In the preceding section, we have figured out what e-properties are. In section 1.1, we got ourselves a simple and plausible conjecture about how to find out causes in the world: We try to see whether a property makes a difference or not, whether intervening with it changes anything in the world. Now it is time to see whether e-properties satisfy our criteria to be bestowers of causal powers. To do this, I will make use of a simple method, which I wish to simply call "the method of comparison", defined as follows:

The Method of Comparison: An object O1 with an e-property P and which causes effect-type E is compared with an object O2 with all the properties that O1 has *except P*. If O2 can also produce E, this shows that an object's *having P does not make a difference* to the occurrence or non-occurrence of E. To put it another way, P does not bestow causal powers on O1 in causing E.

This is nothing more, nothing less than the scientific trial-and-error method. Causes are normally found out by experimentation. Here, we will do thought experimentation. As it is practically hard for us to get such ideal cases where two objects share all their properties except one, especially when organisms are in question, we will imagine ideal objects and ideal circumstances. I believe I have justified the use of this method in section 1.1.

The "plot and characters" of the thought experiments will be standard settings and creatures used in many twin-involving thought experiments about mental causation and mental content, so this will also help us link the current discussion with the discussions in the literature.

2.2.1 Causality Check for Historical E-Properties

We will begin with checking the causal efficacy of historical e-properties.

The character we will use as our "control group" is Swampman. We met him before in section 2.1, but let's reintroduce him to be clearer about whom we are talking about:

Swampman: An entity which suddenly comes into existence as the result of a chance event, and it is identical, in all its properties, to a person which gets destroyed while Swampman comes into existence, except for his historical e-properties. He has a very different history than the person who was destroyed. (I will refer to the destroyed person as "Man".)[42]

Now, let us imagine that I am destroyed, while writing these sentences. An ultra rare chance event happens and in my place, there appears another entity. We can think of it as a "microphysical" duplicate, however, the way we defined Swampman above is silent about physicalism. He has everything I have, and if I had any non-physical parts, he has them too. What is important is that he doesn't have the history that I have. His brain states have never contributed to the survival or the propagation of an organism. They have never been caused or triggered by things in his environment. Historical externalist theories have no way of attributing mental states to my swamp twin.[43]

[42] Why Swampman is named as such is that when Davidson came up with it in his 1987, the story involved a lightning striking a swamp which leads to, by pure coincidence, the emergence of Swampman (while another lightning destroys Davidson). The flavor is, naturally, inessential for the formulation, and one can think of the emergence of Swampman any way one wants as long as Swampman is denied the possession of e-properties. It also seems inessential for Man to be destroyed, but setting the scene that way sometimes eases the exposition.

[43] Let me make a short clarification about the methodology. I will not be using Swampman-like thought experiments as "intuition pumps" (a term due to Dennett 1984). That is, the Swampman thought experiment will not be used to say to the reader "come on, this guy behaves just like you do, of course he has got mental states!" On the contrary, for the sake of the argument, I will grant the strongest form of externalism, and assume

Let’s say that when I was destroyed, I had certain mental states: I had a desire to smoke a cigarette, I had beliefs about where my cigarettes are, and how a lighter should be used to light up a cigarette. Let’s say that if I were not destroyed, it was determined that I would light one up. Lighting up a cigarette might be a mundane piece of behavior, but from a scientific standpoint, it is a rather complicated one which is interesting to investigate the mechanisms of. We ordinarily assume that one needs to have a lot of discriminatory, conceptual, perceptual and motor abilities to light one up. Swampman has none of these, according to historical externalism. Common-sensically, someone would explain my behavior by citing these mental states, saying that I lighted up a cigarette because I desired to smoke one, not because, for instance, I had a desire to pretend like I am smoking, or as opposed to my hand being moved by someone else. How I could successfully do it would also be explained by citing my beliefs about cigarettes and lighters, and my perceptual states. Sophisticated scientific accounts, mentalistic ones, would also attribute many sub-personal mental states to me, maybe with an aim of explaining my behavior in a more detailed way. Whatever the behavior of Swampman will be, it can’t be explained through these if historical externalism is true, because he lacks mental states.

Now, we are at the notorious time t, the time that Swampman comes into existence. What will happen at t2, right after he comes into existence?

that creatures like Swampman *don’t* have any mental states, so that I can compare them with their minded twins to establish the causal inefficacy of e-properties. Of course, if our purpose were to test intuitions or establish a conceptual analysis, then it would be totally legitimate to use Swampman-like thought experiments that way.

More particularly: Will anything different happen, will my swamp twin fail to be the cause of anything that I could cause?

My swamp twin has exactly the same bodily state as me, including the states of his brain. He has the same kind of neural structures and the same firing patterns. From what we know about physics and chemistry, we expect him to go through the same bodily changes I would if I were not destroyed. After all, the same firing patterns will result in the same spread of activation, the release of the same neurotransmitters, which will culminate in the same bodily movements. His hands will end up on the cigarette and the lighter, and at the end the cigarette will be lit up. He will have the same effects on the world, the world including his body. He will do all this *without* having the e-properties that I had. Even if he had any, they wouldn't make a difference. If he goes to an "externalist rehabilitation center", if he gets all the causal connections between the world and his brain states, that wouldn't change anything as long as he had all the non-e-properties right. That wouldn't add anything to his causal powers.

I take this to be a rather uncontroversial assumption. After all, we would, scientifically and common-sensically, take it to be really weird if Swampman's intrinsic physical states led to different behavior or no behavior at all. That would go against what we know about how physics, chemistry or biology works.

Me and my swamp twin have lots of things in common, including our causal powers. We don't have any e-properties in common. As another entity can cause the same state of affairs without having those properties, my e-properties do not bestow any additional causal powers on me, therefore, e-properties do not bestow any causal powers on the entities that have them, and do not causally explain what they do.

What does this imply for externalism and mentalistic causal explanation? If historical externalism about the mind is true, an entity that doesn't have a mind can have exactly the same causal powers as one that has a mind, and mentalistic explanation of how and why I smoked a cigarette is bogus, for my mental properties do not bestow any causal powers on me and invoking their existence is not an explanation of how and why I do what I do.[44]

Above, we looked at the matters in the context of pure historical externalism. It is also possible to tell a similar story in the context of a two-factor theory. Some two factor theories would allow Swampman to have some mental states. For instance, a two-factor theory can say that my Swamp Twin has a desire to smoke something which is called a cigarette and which is lit up by something called a lighter, even though his thoughts wouldn't refer to cigarettes and lighters because narrow content by itself is

[44] Some readers might have a worry here. Didn't what I have just told also imply that mental properties do not causally explain behavior *anyway*, regardless of externalism or internalism being true or not, given that mentalistic sciences do not mention any of the physical-chemical phenomena I have mentioned above, which seem to be the real causes anyway? No. First of all, for the purposes of this thesis we are totally silent about matters regarding physicalism or anti-physicalism. According to an internalist story, mental properties might be just those physical properties we have cited, even though this possibility today is not regarded as plausible in many circles, physicalist or non-physicalist, for reasons related to the multiple realizability of mental states. However, even if this near-consensus is right, note that internalistically understood mental properties can still *synchronically co-vary* with the above mentioned physical-chemical properties, which may bring back their causal relevance, if not causal efficacy. If there is a lawful correlation between those physical states and mental properties, then it is not possible to run a contrastive thought experiment like the one above.

not enough to fix the reference without certain historical facts. For a two-factor theory like this, our thought experimentation would follow a similar course, showing that it is only the shared narrow content which is causally efficacious.

It is also possible to establish the causal inefficacy of historical e-properties in an Earth-and-Twin Earth setting. We have already been there when going over Putnam's now-classical thought experiment. (Putnam 1975) For the record, let's see them again:

Earth & Twin Earth: A possible world where there are two planets, one being identical to our Earth, the other one being identical to our Earth except for one respect: Wherever there is H_2O on Earth, there is a liquid which we will dub XYZ on Twin Earth. It's a time before modern chemistry, and H_20 and XYZ share numerous properties such that their difference cannot be told without the methods of modern chemistry. People on both planets are identical in all their properties except their relations to these liquids. For people on Earth, H_2O is in the causal chain leading to their brain states that subsume their utterances of the word form "water" and certain beliefs and desires involving the concept they express through this word form; for people on Twin Earth, it is XYZ.

In his 1997, Paul Boghossian came up with a third planet for this possible world, *Dry Earth*, which we will also borrow for our purposes. Its name looks descriptive enough, but let's be clear:

Dry Earth: A planet which is identical to our Earth except for one respect: There is no H_2O on Twin Earth, and unlike Putnam's Twin Earth, nor is there another liquid on Dry Earth that shares the same superficial properties with H_2O. It's a time before modern chemistry. People on both planets are identical in all their properties except their relations to H_2O, since there is no H_2O on Dry Earth, and H_2O has never been in the causal chain leading to the brain states of Dry Earthians that subsume their utterances of the word form "water" and certain beliefs and desires involving the concept they express by using this word form (if they express any).[45] When dry Earthians say to themselves "Let's imagine water" and do it, they are in the same phenomenal state as us when we imagine water. [46] [47]

First, let us think about what a person on Earth can do. Someone on Earth, just like someone on Twin Earth, believes that there is a liquid called "water" which quenches thirst, which flows from taps and which makes

[45] Some radical externalist theories do not allow Dry Earthian "water" utterings to express any concept at all. See especially Millikan 2004-a and 2004-b.

[46] I am not sure if Boghossian would define Dry Earth in exactly this way. My aim is not historical precision, but adopting a useful idea for my current purposes. Also note that Boghossian had a different purpose in coming up with Dry Earth: He wanted to give a *reductio* of externalism, trying to show that if externalism is true, then we can a priori know that *whatever we can think about exists*, which is absurd. The current aim is not a falsification of the externalist theory outright, but establishing its incompatibility with mentalistic causal explanation.

[47] There are a few well known complications with these thought experiments, but I take it that it is agreed upon that they do not matter. For instance, Putnam tells us that people on Earth and Twin Earth are intrinsically alike. But actually the human body contains a lot of H_2O, so people on Twin Earth are very different from us in this respect. This doesn't matter because no internalist believes that mentality depends on H_2O in the body, and if someone did, then the H_2O example could be changed to some other natural kind not found in the human body. Similar considerations apply to Dry Earth, and the reader may consider that people there have some stuff in their body that fills the biological role of water without sharing its other properties. It is also assumed that until the discovery of the chemical structure of water, things unfold the same way in both worlds. But it can't: If H_2O and XYZ are distinguishable by science, then they need to have some different effects somewhere. This is not a problem, since those micro-effects of H_2O and XYZ need not matter for our purposes. However, examining these complications may still reveal some background assumptions that lead to some methodological flaws. See Farkas 2003 and 2008 for a criticism of Putnam's thought experiment which is reached through reflecting on the first complication mentioned above.

you wet when it is falling from the sky, etc. We expect him to do things like going to the sink and opening up the tap when she is thirsty, or take an umbrella when the weather forecast says that it is raining. It looks like what they share, either their narrow content or their microphysical properties, is enough to explain their behavior, and their different wide content doesn’t make a difference. If presented with some XYZ, Earthians would reach out for it if they were thirsty, just like Twin Earthians do.

At this point, things get a bit more complicated than the Swampman case above. First of all, in the Swampman case, one of the entities we used in our comparison lacked e-properties. Here, both of our subjects have e-properties. It might be possible that both e-properties are causally efficacious. People on Earth reach for H_2O, and also for XYZ if they go to Twin Earth, in virtue of their desire for H_2O, and people on Twin Earth reach for H_2O and XYZ in virtue of their desire for XYZ. It can be the case that *both* e-properties are causally efficacious; they just bestow the very same causal powers. Although this line of thought may sound appealing to some, we should remember that the people on Earth and Dry Earth share the same narrow content. If we can explain their behavior through the narrow content, then we can see that the wide one doesn’t make a difference.

Another complication is that what we have termed “narrow content” above might be also be wide after all, even though it is “narrower”: The mental properties people on Earth and Twin Earth share, like having a belief that there is a liquid called “water” that flows from taps, might depend on their causal relations to taps and the word form “water.” So e-properties can still be causally efficacious.[48]

To overcome these complications, let us think about a person on Dry Earth. Even though there is no H_2O or XYZ on Dry Earth, this person has the same narrow beliefs about the world like those on Earth and Twin Earth. He believes that there is a liquid called “water” which flows from taps, quenches thirst, and falls from the sky. We expect someone who believes in those things to do what people on Earth and Twin Earth do: To look for a tap when she’s thirsty, to take an umbrella when she hears the sentence-form “it is going to rain”, etc, even though her brain states don’t have properties like “being caused by H_2O” or “being caused by XYZ.” She would do the very same things if she had those e-properties. If Dry Earthians don’t need these properties to have the dispositions they have, it looks like for the case of Earthians and Twin Earthians, it is not their e-properties that bestow causal powers on them.

Here too, we should think about the possibility that the so-called “narrow content” of Dry Earthians may also be wide after all, that it depends on their relations to taps and other liquids. But to block this complication, we can think of Dry Earthians as living in a world where there are no taps or “water” utterances at all. However, the externalist can still push us towards thinking that these thoughts too can be wide: Dry Earthians’ belief that there are things called “taps” may depend on their relations to the sounds “t”, “a”, “p” and “s”, their relations to certain material and shapes that can be “put together in their minds” to form tap-thoughts etc. We could

[48] This point can be found in Jacob (1997).

probably keep playing this game until we push the externalist to a corner where her externalism starts to sound rather unintuitive as the number of causally relevant mental states gets fewer and fewer. However, if we do not want to regress into an almost solipsistic world to establish the causal inefficacy of historical e-properties this way, it is better to resort back to a Swamp Dry Earthian, or just a Swampman on Earth.

2.2.2 Reflections on the Causal Inefficacy of Historical E-Properties

We have seen that historical e-properties fail to pass our test for being the bestowers of causal powers, if we are on the right way in thinking about causation and if we haven't taken a misstep while conducting our inquiry. Showing that e-properties are causally inefficacious does not tell us *why* they are inefficacious. In this section, I will shortly reflect on why we expect me and my swamp twin to have the same causal powers. This is not necessary to establish my argument, but hopefully it will lead to a better understanding of how we expect causation to work. First I will point to some examples where many people, including externalists, either implicitly or explicitly take historical e-properties to be causally inefficacious, and then I will try to see *why* we take them to be so.

Me and my swamp twin having the same causal powers is intuitive, so intuitive that even externalists seem to assume it. For instance, when setting up Swampman-style thought experiments, it is generally assumed during the set-up that Swampman and Man are *behaviorally indistinguishable*, that they cannot be told apart by observing their behavior, which means that they have the same causal powers. I refrained from saying that in my own setting up of the thought experiment, for that was just what we wanted to find out, not something we would like to assume from the beginning. But it is generally assumed, and this is a sign of a widespread intuition. Here is a passage from Dretske, where he seems to surrender to epiphenomenalism about the mind while building his teleosemantic externalistic account:

> Not only does the mind not supervene on the current physical state of a system, it does not supervene on the current global state of the universe. According to teleosemantics, what we think and experience today – indeed, the fact that we think and experience anything at all today – depends not only on what is going on in us and around us, but on events and conditions that existed long ago and (probably) far away. A physical duplicate of a conscious being, a person (?) who lacked the appropriate history – a history that gave its internal states the requisite functions – would not think and experience anything at all. *The internal machinery would function – causally speaking – in the same way*, but it wouldn't have the same (or, indeed, any) function. (Dretske 2006, p.75, emphasis added.)

This point is not specific to Dretske. The causal inefficacy of historical e-properties seems to be inherent to use-theories of content. Use-theories tell us that the content of your mental states is fixed *after* you do something. For instance, only after a brain state of yours (or of one of your ancestors) helps you (or one of your ancestors) behave adaptively towards snakes that it acquires the content SNAKE. But this implies that you were fully able to behave adaptively towards snakes *before* you had a state with the content SNAKE, and having the concept SNAKE or thoughts about snakes does not add anything to your adaptive behavior.

Having a historical e-property is standing in a relation to something that existed in the past. That thing that existed in the past does not exist now, “now” being the time that we are trying to see what kind of causal powers something has. It is generally taken for granted that what determines the causal powers of something now are the things that exist now. It is hard to see how the properties had by the snake that caused a brain state I have now can bestow any causal powers on me now. It could have been the case that the same brain state, non-historically understood, could be caused by a garden hose reflecting the same kind of light, and what I would have now would be those non-historical properties of the brain state that gives me my abilities. It is hard to see how the fact that my brain state was caused by a snake, a toy snake, or a garden hose would make a difference to my behavior *as long as it is the same state that is caused in me*, which I have now. And even if things that existed in the past but don’t exist anymore could have effects on what will happen now, it would be unclear in which sense what I do now is a matter of my causal powers or abilities, rather than a matter of the causal powers of the thing that existed in the past.

It is not only hard to see. This kind of thinking is everywhere in our dealings and manipulation of the world. If I need an ashtray, an object that has the capacity to hold ashes inside, I can make an ashtray, or buy something that was designed as an ashtray, but equally, I can go out and find an ashtray shaped piece of wood, or use a cup which was designed to hold liquids inside. The historical properties like “being designed as an ashtray”, “having the proper function of holding ashes inside”, “having the proper function of holding liquids inside”, “being designed by Hendrik” or “being caused by a natural storm" seem to be wholly irrelevant in what that object can do as long as it is made up of the same material and has the same shape, and it is these properties that explains how it can hold ashes inside. Similarly, when we are trying to manipulate the world or trying to see how things work, we manipulate or look at the properties that exist now. We cannot do otherwise, anyway. We can manipulate the material and the shape of the ashtray-shaped object, but we cannot manipulate the events that gave rise to its existence, for they are gone now. We may never know how it came into existence at all. The important thing is that *it doesn't matter*. We understand how things work and we manipulate the causal powers of objects without knowing how they came into existence.

Some may think that the above reasoning is too quick. Knowing what something is caused by can sometimes aid us in finding out what it can do now. For instance, if we know that something is designed for holding ashes inside, there is a good chance that it has the disposition to hold ashes inside. However, this does not mean that the historical property bestows any causal powers on the ashtray; such knowledge can only get us closer to the properties that actually bestow causal powers. Something doesn't hold ashes inside at time t in virtue of being designed that way, but in virtue of the properties like shape, size and material that designers had put together. Anyway, having such historical properties doesn't guarantee the existence of causal powers. A broken ashtray or one that is made out of the wrong material cannot hold ashes inside even though it has the property "being designed to hold ashes inside." Going from historical properties to causally efficacious properties is a matter of luck.

What I want to ask, before passing, is that whether the converse is even imaginable, whether it is possible to imagine a reality where, for instance, the ashtray-shaped piece of wood I find on the street doesn't hold ashes inside *just because* it wasn't designed to hold ashes inside, or that my swamp twin cannot adaptively deal with snakes *just because* his internal states were not caused by snakes, even though he has all the machinery inside. My intuitions about the coherent imaginability of such a scenario are mixed. However, even trying to imagine such a world shows us better that our world is *not* like that. The above described situations sound almost magical. And for all we know, the way we deal with the world, knowing and manipulating the non-historical properties of things, simply works.[49]

Before passing, let me touch upon on the implications to the philosophy of artificial intelligence these considerations about the causal inefficacy of historical properties have. One of the replies given to Searle's notorious Chinese Room Argument was the "robot reply". (Searle 1980) The robot reply granted that the Chinese Room that Searle imagined really didn't have any mental properties, but it would acquire them if the room was a part of a robot that interacted with the world. But what seems to be wrong with this reply is that the Robo-Chinese Room, or any other robot, is

[49] These considerations might even push one towards the idea that there is *no such thing* as a historical-relational property, for one cannot stand in a relation to things that don't exist anymore. Within the confines of this thesis, I don't want to, and don't need to, go that far. It is also not very clear what we would say for such a world as described above. A world where my swamp twin cannot adaptively deal with snakes can be a home to both of the following possibilities: The historical properties might have causal efficacy, or my swamp twin lacks some undetectable non-historical properties, non-historical properties that can only be caused by snakes. For the second possibility, my adaptively dealing with snakes wouldn't be in virtue of my brain states' having the historical property of "being caused by snakes" or "having the proper function of indicating snakes", but in virtue of some non-historical properties that can only be caused by snakes and that are otherwise not detectable. A third possibility is that in such a world, causation doesn't work in a temporally local way, and the snake that existed in the past is exerting some causal force on what I am doing now, if such a possibility is not contradictory to our coherent understanding of time and existence. In such a case, what I do would be partly a matter of the causal efficacy of the properties of the snake in the past, that is, it wouldn't be a manifestation of the causal efficacy of my relational property, but the snake's intrinsic properties.

already disposed to do what they can do before they interact with the world. The early claims about the possibility of artificial intelligence rested on Turingian ideas about behavioral indistinguishability, which conflict with denying mental properties to machines that do not fulfill historical-externalistic criteria. If the philosophy of artificial intelligence were to be historical-externalistic, it would be a refusal to attribute mental properties to computers that came fresh out of the factory, with the "software" installed, which many defenders of artificial intelligence would find absurd.

2.2.3 Causality Check for Non-Historical E-Properties

Now we have established the causal inefficacy of historical e-properties, although we are yet to face the objections, and I have suggested that this might be due to the fact that historical relational properties are relations to things that don't exist anymore, and things that don't exist anymore can't be causally efficacious now. And if they were really efficacious, then it would be a matter of their causal powers, not of the entity that is standing on the other side of the relation. Now it is time to see whether non-historical e-properties pass the test.[50]

[50] Let me put forward a reservation about purely non-historical externalist theories. It seems to me that no theory of mental content can be purely externalistic and purely non-historical at the same time. Synchronic externalist theories give us e-properties like "being in the same world as X." However, *all* states, events and objects in the world have that property, and it seems like we need to point to some special property of organisms or their brains. As synchronic externalists do not offer us another relational property, this is likely to be either something other than an e-property, which will turn the theory into a two-factor theory rather than a purely externalistic one, or the externalist should offer us a historical property. Accordingly, I will assume that all non-historical externalist theories

To test the causal efficacy of synchronic e-properties, like “being in a world where there is H_2O”, we will use the Earth-Dry Earth setting again. However, this time we should describe the setting in a way that Earth and Dry Earth are not two planets in the same possible world. For if they were, then people on Dry Earth would also have the property “being in a world where there is H_2O”, for there would be H_2O in their world (“world” understood as “actuality”), even though it is not on their home planet. For this reason, we will imagine Earth and Dry Earth as existing in different possible worlds.

Let us think about something that Earthians are disposed to do in their world: They go and try to open a tap when they are thirsty. As we already now, Dry Earthians also believe that there is a liquid called “water” which flows from taps, and for all we know, they would be also disposed to go and check the taps when they feel thirsty, just like people here on Earth go and check the Atlantic Ocean to find Atlantis. It looks like being in a world where there is H_2O does not bestow any causal powers or dispositions on people, or being in a world where there is Atlantis, in that case.

Actually, we may not even need another possible world to run our thought experiment. We can think about two different times slices of Earth, when something exist and something doesn’t. For instance, according to a synchronic externalism, people cannot have thoughts about mammoths after mammoths cease to exist. Nevertheless, if a caveman has a narrow belief that there are some creatures called “mammoths” that he can hunt

are two-factor theories, acknowledging the existence of certain mental properties not being dependent on or identical to e-properties.

and eat, we expect him to get his weapons and go out, just like he did the day before when he hunted the last mammoth on Earth. Of course, an externalist would want to define her theory in a way that “being in a world where there are mammoths” would apply to all time slices of the world, building up a non-causal non-synchronic externalism. In that case, we can resort to Dry Earth or Atlantis, where the entity doesn’t exist in any time slice.

The above examples are provided assuming that all non-historical externalist theories are two-factor theories. For pure externalist non-historical theories out there, if there are any, we wouldn’t be able to rely on the narrow content to make our point. But in that case, we could still rely on the brain state of the mammoth hunter before and after the extinction of mammoths, which would force us to think that the hunter would behave in the same way, as we have seen with Swampman.

2.2.4 Reflections on the Causal Inefficacy of Synchronic E-Properties

When reflecting on the causal inefficacy of historical properties, I have suggested that possibly, these properties never bestow any powers on anything, and if we are able to successfully imagine a world where they do, we find a rather weird world and we cannot really decide what to say about it. This might not be the same for many synchronic relational properties.

It is possible to come across remarks to the effect that relational properties are, in general, causally inefficacious, that things can only affect each other’s causal powers by affecting their intrinsic properties. However, this might not be the case, for modern physics seem to suggest the existence of certain relational properties which might have causal efficacy. An object’s being in a gravitational zone is such a candidate property. Here, it looks like standing in a relation to another object bestows causal powers on the object at the other side of the relation. Two objects with the very same intrinsic properties may differ in their behavior depending on the gravitational relations they have. “Being entangled with something” might be another candidate, but I do not wish to consider it in detail for I lack the competence to do so.

The first thing to say about such power bestowing relational properties is that they are rare and unintuitive. Gravity was named an “occult” force when it was first introduced, for an object seemed to determine the fate of another without a spatiotemporal causal chain leading from the first object to the second that changes the second object’s intrinsic properties. Quantum phenomena were similarly called “spooky action at a distance” by Einstein. I take this to be data for our regarding such power bestowing properties as unintuitive, contrary to the way we generally think about the world, even though newer generations wouldn’t go as far as thinking of them as “occult” or “spooky”. Nevertheless, it seems that there were occasions where we had to postulate such power bestowing properties to understand how things work, unlike in the case of historical properties.

Would we ever need to commit ourselves to the causal efficacy of e-properties pointed out by synchronic externalism? This question, again, suggests us to imagine a reality where those properties are causally relevant, a world where we would need to postulate a power bestowing nature to those properties to explain things that would otherwise go unexplained.

How would such a world be? Think about our water-thoughts and water-utterances. Then imagine that all the H_2O in the world dries out and Earth turns into a Dry Earth. In such a case, we would attribute some causal efficacy to the property “being in the same world with H_2O” if we would suddenly lose some of our capacities, if we stop manifesting some of the behavior we manifested before. Imagine a world where, *at the very moment* H_2O ceases to exist, people stop uttering the word “water”, stop asking for “water” when they are thirsty, stop going to the lavatory, lose the capacity to hypothesize about water and chemically produce it in labs, etc. Imagine that all of this is happening without any change in the intrinsic properties of those people.

This scenario is quite implausible, and its implausibility varies according to how externalistic we are about mental states. If someone were to preserve the same physical-chemical states in one’s body but would stop manifesting a particular behavior when the last drop of water dries out, that would be quite “occult”. The case seems even more implausible if we go for a two-factor theory, for that person not only would have the same states defined in a physical-chemical way, she would also have the narrow beliefs associated to “water” thoughts, such as there is something called “water” and it quenches thirst.

For sure, someone could say that the “magical” impression of this phenomenon would fade away as it did for gravity, which could be right. But nevertheless, is it plausible to think that we will encounter any such phenomena in our world to make us believe that such e-properties are causally efficacious? For all we know, people *do* exhibit various kinds of behavior even when the referents of their narrow states don't exist, like going to the Atlantic Ocean to find Atlantis. And for all we know, this behavior is exactly the same as one that would happen if Atlantis did actually exist. For sure, strong versions of externalism deny that we have Atlantis-thoughts. However, we can make our point by pointing to the neural properties of people who end up in the Atlantic Ocean, if one thinks that Atlantis-thoughts do not exist.

Therefore, it seems wildly implausible to think that we will ever need to attribute causal efficacy to these properties. But a question seems to remain: What is it about these properties that makes it implausible to attribute causal efficacy to them while it is not implausible for gravitational relations? Let us see. The first difference is that gravitational relations are something we discover, "my being in the same world with H_2O" is not, in the sense that such a relational property follows logically from the existence of me existing and H_2O existing. If I know that I exist in this world, and if I know that H_2O exists in this world, it logically follows that I am standing in a relation to H_2O, a relation of being in the same actuality. However, the relation of this ball being in the gravitational field of that ball doesn't logically follow from any non-relational properties, at least not in the same way that is relevant to our discussion here.

But why does this difference matter for causality? Here is what I tentatively propose: If a relational property P logically follows from properties Q and R, and if there is some causal work being done, we are tempted to think that all the work is done by Q and R. For if we know that Q and R are causally efficacious by themselves (even when they don't exist together), and if we can causally explain what is happening by relying on these properties, there is no reason to think that the relational property makes a difference. If I can cause something by the non-relational properties and if that snake over there can also cause something with the non-relational properties it has, and if what we cause when we are together is just what we cause alone when we could cause separately added to each other without anything genuinely new, then there is no reason to think that the relational property that logically follows adds anything to matters causal. Things are *not* like this for gravity and entanglement.

2.3 Properly Setting Up the Argument

In the preceding sections, we made an attempt to see whether e-properties are causally efficacious or not, and concluded negatively. As a bonus, we reflected on why these properties are inefficacious which led us to some tentative but plausible grounds to support what we have found. Now it is time to take a step back and look at the bigger picture, to see where we can go from here. Could it be that we have gone wrong in a step of our reasoning? If not, does this imply that externalism has no place in mentalistic sciences or are there other, non-causal doors open for externalism? Does all this falsify externalism or vindicate internalism, or do we have other choices?

After having showed that e-properties are causally inefficacious, we will try to see what may or may not happen to externalism after it concedes to this fact, or whether externalism can resist it or not.

Let's remember what an externalist theory which claims to be the backbone of a mentalistic causal-explanatory science is committed to. It is committed to the following three statements:

(1) Mental properties are real.

(2) At least some mental properties bestow causal powers on the entities that have them.

(3) All mental properties are e-properties.

However, if we introduce what we have found into the picture:

(4) E-Properties do not bestow causal powers on the entities that have them.

we end up with a contradiction which needs to be resolved:

(5) If (3) and (4) are true, (2) is false.

which means that *a causal-explanatory science of behavior that adopts an externalistic account of mental properties has a logically incoherent theoretical framework.*

This is where we are now, or to put it better, that is where externalism is now. It is faced with a contradiction to be resolved. The ultimate aim of this thesis is to indirectly contribute to a refutation of externalism, by showing that it is (3) which needs to be abandoned. But we first need to see whether externalism can get out of this contradiction without abandoning (3).

A consistent externalism needs to show the falsity of either (4) or (5). To refute (4), one needs a demonstration of the causal efficacy of e-properties, an argument to the effect that there is something wrong in the way I thought I had established their causal inefficacy. To refute (5), one needs to show that (2) and (4) are not incompatible and the word “cause” is used in different senses in the two statements. That would show that, contrary to what I claim, there is no real contradiction here. Nevertheless, I will not consider this option here, for I cannot think of what the two senses of the term might be.

If no plausible objection can be given to (4) or (5), externalism has to do without (2), making it an epiphenomenalist theory, and the externalist project of providing a constitutive analysis of mental states for a causal-explanatory science of behavior will fail. If epiphenomenalism is unpalatable, (3), externalism itself, should be abandoned. From there, the convert has two ways to go: The way of eliminativism by denying (1), throwing the baby out with the bathwater, or the way of internalism by affirming a positive internalist statement.[51]

2.4 Giving Externalism a Chance and Blocking the Objections

We will begin with objections to my claim (4), that e-properties are not the bestowers of causal powers. An objector to that claim has to show us that after all, e-properties do bestow causal powers. What could be wrong in the way that I have tried to establish the causal inefficacy of e-properties?

[51] What if a mental property is a conjunction of an e-property and some power-bestowing property which is not an e-property? Will that save externalism from the causal challenge? No, for it won’t be the mental property, but only a part of the conjunct that is bestowing the causal power, and causal explanations would be possible without mentioning the e-property. Thanks to Ceyhan Temürcü for pushing me on this point.

If there is something wrong, it is probably *not* due to my mistakenly thinking that Swampman or Dry Earthians can do this and that. Everybody seems to agree that if I am replaced by my swamp twin when I am heading to the library, my twin will also end up in the library. Nobody in the literature thinks that a human organism created in the lab which is intrinsically identical to a human gardener wouldn't have the ability to cut grass. And, nobody thinks that people on Dry Earth won't go to (dry) taps and fountains when they feel thirsty.

So, if I did something wrong, that is probably my *forgetting* something rather than being mistaken about the common powers of Man and Swampman, my forgetting to consider something that Man can do but Swampman can't, something that we can do but Dry Earthians can't. What could that be?

We will now look at some examples from the literature which suggest what that might be, what might be the things that entities with e-properties can do but those without can't. But first let's see an objection which suggests that I forgot a more general fact about causation: Overdetermination.

2.4.1 The Overdetermination Response

Overdetermination is a term of art in the metaphysics of causation. Cases of overdetermination are cases where we have two things or two properties both bringing about the same effect. A commonly suggested example is the example of two bullets fatally hitting someone at the very same time. The bullets overdetermine the person's death. Both are individually sufficient for her death, and both are causally efficacious.

It is a matter of debate whether we have genuine cases of overdetermination in our world. For cases similar to the above example that happen in our world, we expect there to be a temporal difference in the hitting of the two bullets, no matter how small, so only one bullet qualifies as the cause.

One could try to block my above argument relying on overdetermination. A defender of externalism could be tempted to say that even though non-mental properties are sufficient for what Swampman does, in the case of Man there are two sufficient causes: Mental and non-mental properties.

I would like to block this kind of response by two means. Firstly, the over-determination response seems to be too cheap, cheap in the sense that it comes too easily. It is possible to argue for the causal efficacy of *anything* relying on overdetermination. Imagine that we are trying to find out whether schizophrenia is caused by chemical imbalance in the brain or evil spirits that possess the schizophrenic person. After we find out that the chemical imbalance is a sufficient cause, there is no motivation left for believing that the evil spirits play any role in schizophrenia. When we understand the neural basis of schizophrenia, we understand that one doesn’t need to be possessed to be schizophrenic, and anyway, being possessed wouldn’t make a difference to one’s schizophrenia as long as one had the chemical imbalance. But relying on overdetermination, one could keep insisting on the existence of evil spirits and their causal role, claiming that the *evil spirits overdetermine schizophrenic behavior*.

I hope the above example is illustrative in showing that overdetermination is contrary to the way we think about and find out causation, and relying on it seems to be very ad hoc. In the real cases that come close to being cases of overdetermination, like the two-bullet case, things that overdetermine other things have effects other than what they overdetermine. For instance, the bullets have the effect of displacing air, other than their fatal effect. In the above case of schizophrenia, when we understand the real cause of the effect we were investigating, we have no reason to postulate evil spirits anymore, for unlike the bullet case, we don't see any other effects of the spirits anywhere else. Same considerations apply to e-properties.

The overdetermination strategist would say that things are a little bit different for e-properties, and she would be right. E-properties are *not* initially postulated to explain anything. Rather, they are properties that we can logically reach: If X caused Y at t1, and if Y still exists at t2, then Y has the e-property "being caused by X once upon a time" at t2. As they logically follow, we cannot stop believing in e-properties on causal grounds. It is impossible to stop believing in the existence of the e-property "being caused by X once upon a time" that Y has at t2 if we know that there was an event of X causing Y once upon a time. However, the overdetermination strategist cannot rely on this fact to establish that e-properties overdetermine behavior, for that would imply that if we cannot stop believing in the existence of something for logical reasons, then that thing is causally efficacious. There is no reason to hold that such a logical fact would imply a causal fact, and the overdetermination strategist should say something extra to make us believe that it does. For one, I don't know what can be said.

This was the first reason for us to doubt the overdetermination strategy. The second one might be more important. In cases of overdetermination, both properties are said to be sufficient for the effect. However, e-properties by themselves are not sufficient for the effect. Every time there is an e-property, there is something that has the e-property, and that something has certain non e-properties which, as we have seen, are sufficient for the effect. E-properties do not join forces with the non-e-properties to bring about the effect. As we have seen, the non-e-properties are enough.

E-properties do not determine anything either. Saying that something has an e-property of "being selected for subsuming adaptive behavior towards snakes" does not tell us anything about what that thing can do unless we know *what* is that thing that is selected. It is pretty well possible that the thing that is selected, which has a historically acquired biological function, is malfunctioning now because of having or failing to have certain non-e-properties. A brain state or a narrow content's being caused by H_2O doesn't tell us anything about what it can do. It is pretty well possible that one has a state caused by H_2O while thinking that water is a species of elephant living in Myanmar, and her behavior following her water-thoughts would be wildly different than someone who believes that water is a potable liquid.

As we have also seen in our reflections in section 2.2.2 and 2.2.4, it is hard to see how e-properties determine things, let alone overdetermining them. Shortly, there is no reason whatsoever to think that the overdetermination response could save the causal efficacy of e-properties.

2.4.2 "People on Dry Earth Don't Have the Power to Drink H_2O": Version 1

Now, we will consider two versions of an objection, with two different settings. In the first one, we have to think about Earth, Twin Earth and Dry Earth, but think of them as being in different worlds, in different realities. When we think of Dry Earth, we have to think of it as, a planet in a world where there is no H_2O or XYZ at all. In the second, we have to think of the three planets as being in the same world, like three planets in a universe.

Millikan, in her 2004-a, seems to suggest that having a capacity necessitates the existence of the thing the capacity is directed at, when she claims that Angel-thoughts do not confer any abilities on those who have it, as she things that mental states confers abilities on their possessors in identifying things, and as there are no angels to identify.[52]

With a similar reasoning, one might be inclined to say that the person on Dry Earth doesn't have certain powers the people on Earth and Twin Earth have. For instance, he cannot drink H_2O because there is no H_2O on Dry Earth, and nowhere else in that possible world. Therefore, my property "being in the same world with H_2O" bestows causal powers on me. As soon as the last drop of H_2O vanishes from the universe, I lose these powers, without any changes in my non-relational states.

[52] Strictly speaking, Millikan doesn't believe that there are any Angel-thoughts, as it is impossible for us to have such thoughts according to her teleosemantic theory. Rather, she calls such thoughts "conceptions". (See her 2004-a)

As far as my intuitions go, this is a strange way to think about one's abilities. Even though I guess most readers would share intuitions with me, some people, such as Millikan, will not, and I have to do something more to ground these intuitions. I will try to ground them in several ways.

The way we generally understand ontological categories like capacities, abilities, dispositions and powers is such that they exist even when they are not exercised. Actually, some of those dispositions, like being explosive, exist only *before* they are exercised. We understand these categories through possible situations. We say that something is soluble in water if such and such would happen if we *would* drop it in water. We say that somebody is inclined to commit suicide if she *would* kill herself in certain situations. Even for those dispositions and powers that persist through their manifestation, they do not depend on being manifested. We say that a person has the ability to solve a problem if she could be able to solve it when she would come across the problem. Before we present her the problem, it is totally sensible to say that she has the ability to solve the problem.

Millikan would not have any disagreements up to this point. But what we have said above can be extended to cases where the thing that is the manifestation partner of a disposition doesn't exist at all. Imagine that all the pianos in the world are destroyed and after a few days a piano is built again. If I have the ability to play a piano, it sounds rather odd to say that I lost my ability to play a piano during those few days.

Thinking about and finding out the capacities whose manifestation partners don't exist is quite common and useful. For example, if we are trying to make a medicine against a microbial disease, it is totally sensible to ask whether the microbes have the capacity to resist the medicine we are trying to create, even though we haven't yet created the medicine, and it doesn't exist. However, the microbes exist now, and now they have the capacity they have, the capacity to resist the medicine we are planning to manufacture.

But there is even more. For some capacities, it is *necessary* that certain things *don't* exist for that capacity to be performed, such as capacities to create, invent or build. The Beatles had the capacity to create the song "I Wanna Hold your Hand" *before* they wrote it, before the song existed. They had the capacity to perform it *before* they first performed it. Their capacities definitely didn't depend on the existence of the song.

However, our Millikanian objector can try to point to this fact: Even though in a world without pianos I have the ability to play a piano, I, strictly speaking, *can't* play a piano. At a time when no piano exists, it is not metaphysically possible for me to play one. No matter what properties I have, I can't play one. This is a matter of what I can and cannot do, and that depends on the existence of pianos.

This is surely right. But it is hard to understand what it has to do with my powers or abilities. It is hard to understand how this general fact about the world, that I cannot play a piano because there are no pianos, is a matter of me and my abilities, rather than being a nomological impossibility *about the world*. The mundane fact that I can't play a piano because there are no pianos, "can't" understood in a strict nomological sense, doesn't change the fact that I *can* play a piano in the sense that I have the ability and I know how to play it.

But there is a bigger challenge for the Millikanian strategy. If my being in a world where there are pianos bestows causal powers on me, then it should causally explain how I can play the piano. It should explain why I couldn't play a piano at time t, when there were no pianos, but I could play at t2, when a piano came into existence. Let's think about such a question, where we demand an explanation of how someone can play the piano, or why someone played a piano. However, we *presuppose*, by asking the question, that there *are* pianos. Therefore, saying "because there are pianos" or "because Sabrina exists in a world where there are pianos" gives us no new information, and no explanation, of how Sabrina can play a piano. And after all, if pianos exist, everything has that same relational property Sabrina has. What is special about Sabrina that makes her able to play the piano while Sabine, or that stick over there, can't? Sabrina, Sabine, and that stick over there all have the property of being in the same world with pianos.

Things may not be the same for the negative question "Why Sabrina *can't* play the piano?" Here citing the non-existence of pianos might give us an explanation, because our question doesn't presuppose the answer, and the answer eliminates other possibilities. But unfortunately, this relational property, Sabrina's being in a world where there are no pianos, is not the kind of relational property we expect to find in an externalist theory, for Sabrina is not standing in a relation to something that exists outside of her. There are no pianos to stand in a relation to, neither she is standing in a relation to non-existent pianos. At best, she's standing in a relation to a *fact*, but it doesn't make sense to say that that fact is internal or external to her. Unlike objects or events, facts are not located anywhere. In any case, she shares relational property with everything else in the world, and the answer doesn't tell us what is special about Sabrina (e.g., that her hands are broken.)

However, things are worse for externalism, for an externalist cannot rely on the above line of response in a coherent way. In a world where pianos do not exist, we cannot have the concept "piano" according to a Millikanian synchronic externalism, and therefore it is impossible to ask the question "Why Sabrina can't play the piano?" and get the answer "Because there are no pianos in her actuality." The word "piano" in these sentences wouldn't express a concept and would be meaningless.

Before passing, let me use this occasion to make a couple of important remarks about Millikan's style of externalism. Many externalist theories somehow begin with the idea that the job of the mind is to *indicate* things that exist, to confer knowledge on subjects about the immediate environment. When we begin with that idea, it is quite inevitable to come to a point, like Millikan, where we will be making the strange claim that there are no thoughts about angels. However, it is far from obvious that indication is the primary function of the mind. Mental properties give us powers in imagining, creating, entertaining counterfactual scenarios, hypothesizing and predicting. It is exactly these sorts of activities, rather than indication, which we are tempted to call "mental". It is particularly when an object is *not* in the immediate environment that we are more tempted to call an ability directed to that object "mental", that we are tempted to say that object exists "only in the mind." Responding adaptively to the immediate environment is what a mindless reflex machine can also do. What a reflex machine cannot do is to imagine the song "I Wanna Hold

Your Hand" and then play it, which also seems to be biologically quite adaptive, given that rock stars have the chance to hold more hands than others.

2.4.3 "People on Dry Earth Don't Have the Power to Drink H_2O": Version 2

Above, we have considered whether I have forgotten about the mundane fact that if there is no H_2O, then, strictly speaking, one can't drink H_2O. We have seen that here the use of "can" and "can't" has nothing to do with causal powers of entities but they point to certain logically obtainable necessities that are totally uninformative, and trying to apply those "can" and "can't" statements to the way we think about powers and capacities has absurd consequences.

Now we will encounter a different version of this objection, which we find rather anonymously stated in the literature. For this objection to be set up, we have to imagine Earth, Twin Earth and Dry Earth as being in the same possible world, like three planets in a universe. Here the objection will not rest on the fact that people on Dry Earth is not in the same world with H_2O. They are. H_2O exists for Dry Earthians too, it is just not found on their planet. The objection rather says that people on Twin Earth and Dry Earth don't have the capacity to drink H_2O because they don't have H_2O in their *environment*.[53]

[53] For an externalist to come up with this objection, her theory should be a non-causal, non-historical one, because she is offering us a non-historical property, "being in the same environment with H_2O", as a causally relevant one. And the theory should claim that the content of "water" thoughts is fixed by what exist in one's environment. However, such an externalism is hard to construct, for it has problems with delineating the borders of those

The general response to this objection is similar to the one we gave to the Millikanian objection above. Causal powers of two entities are to be tested in the *same context*.[54] Think about a man whose feet are tied, and who cannot walk because of this; and a man whose feet are free and who can walk. For the man whose feet are tied, we don't say that he doesn't have the ability or the power to walk, or that he loses this power while his feet are tied, even though we might say that he is "unable" to walk. The free man and the tied man both have the ability to walk; it is only that the ties prevent one of them from walking. He *would* be able to walk if he were in the same context as the untied man. Similarly for people on Dry Earth: They would be able to drink H_2O if they traveled to Earth.

The objector may respond to the above line of thought in this way: "The idea that causal powers should be tested in the same context *presupposes* that synchronic relational properties are causally irrelevant. That line of thought can't prove that synchronic relational properties are causally irrelevant because it presupposes that conclusion. After all, what it means to put two different things in the *same context* is *keeping all their relational properties the same.* But if we make all their relational properties the same, of course we won't be able to see the difference in causal powers bestowed by those relational properties."

"environments". Such an externalist has to tell us something like the following: The environment is the planet, and people on Earth can think about water while people on Dry Earth can't because they don't have it on their planet. However, it is hard to see why anyone would have any motivation to hold such a theory to be true. After all, what is so special about planets, rather than galaxies or towns, that they fix the content of people's thoughts? Even though this kind of externalism is hard to motivate, I have to consider an objection based on it because it is possible to come across claims that implicitly adopt such an externalism, especially in personal communication.

[54] See Fodor 1987 for a defense of this idea.

The objector is right that there is such a presupposition. However, that only shows us something interesting and important about the way we think about the causal powers of individuals. Why we keep the context fixed when we compare the causal powers of me and my Dry Earthian twin is because we are trying to understand the powers *I* have, not the powers me plus my current environment has. As we have said in section 1.2.2, the debate is not a debate about causal powers of my environment and how it can cause a difference in me through affecting my intrinsic properties. But when we also take into consideration the intrinsic properties of H_2O while we're trying to find out my causal powers at any given time, it is unclear in which sense they are the powers of *me*, rather than the causal powers of *me plus H_2O* at any given time, as we have remarked before.

2.4.4 Dual Explanandum Strategies

In the preceding section we have considered an objection which states that me and my twin on Dry Earth do not have the same causal powers because my twin cannot drink water, for there is none in his environment. We refuted the objection by saying that causal powers of two different individuals are to be tested in the same contexts, and my dry twin would drink water if he were here. We justified our basis for testing causal powers in the same context by saying that if we include things that are not a part of me as the basis of my powers, such as H_2O in a bottle in front of me, it doesn't make sense to call it *my* causal powers.

The next thing we will do is to try to see whether the twins I have compared are also doing *different* things *in addition* to what they do in common, something they can do in virtue of the e-properties they have.

2.4.5 The Action Theoretical Strategy

When I tested the causal powers of twins on Earth, Twin Earth and Dry Earth, I claimed that they are doing the same thing. When we put a bottle of H_2O in front of them when they are thirsty, they can, and do, reach out and grab it, for instance. One line of response to the way I tested their powers this way goes by saying that these three people are actually doing different things. According to the causal-informational theory of mental content, even when they have a bottle of H_2O in front of them, they are performing *different actions*: I have a desire for H_2O, and my Twin Earthian twin has a desire for XYZ (even though, as we should remember, we have no idea about the microstructure of what we call “water”.) Accordingly, what I do is *reaching out for H_2O*, what my twin does is *reaching out for XYZ*. We are doing different things, two different things are happening in the two cases. We are acting in a different way.

The same goes for me and my swamp twin. We should remember that according to strict historical externalism, my swamp twin doesn’t have any mental states at all, including desires: When my hand reaches out for the bottle of H_2O, I am performing an action, I reach out for H_2O. My swamp twin is not performing an action at all. He is not reaching out for anything, for he has no desires. Merely, there are some bodily movements going on, his (its?) hand moves towards the bottle. Two different things are happening. Therefore, the objection goes, I was wrong to say that the same thing follows from the state of me and my swamp twin after time t. Therefore, historical e-properties make a difference to what is happening in the world.

This is the *action theoretical strategy* to find the causal efficacy of e-properties. We'll shortly call it ATS. ATS is a species of *dual explanandum strategies*. They are called that way, because the strategies tries to solve the problem of mental causation in a way of showing that mental properties and non-mental properties are brought up to explain different things, and if we ask the right causal questions we can see the causal efficacy of mental properties.

Note that these strategies are not only relied upon in the context of externalism. ATS can also be used to solve the general problem of mental causation, a problem we are not dealing with here. But we are in the context of externalism, so now we shall more clearly see what properties we have here as mental properties, and what they are supposed to explain, according to ATS.

The mental properties we have here are causal-historical properties. My brain states have properties like "being caused by H_2O", "being naturally selected to indicate H_2O" or "having been used by an adaptive way towards H_2O", while my Swamp Twin's brain states have no such properties. My swamp twin, mindless according to historical externalism, can surely do (in the non-agentive sense of doing) a lot of things that I do. He (it?) can keep his bodily integrity for years in dangerous conditions, can dodge the things thrown at him, can "play" chess and win, etc. But he cannot act. The above mentioned properties are supposed to be causally efficacious for his actions. They are supposed to bestow causal powers upon him in acting, and they are supposed to causally explain his actions, according to ATS. Mental properties do not explain the raw physical properties of bodily movements, but some other properties of them.

First we should ask what an action is in the first place. An action, according to the philosophical consensus, is a bodily movement caused by a mental state.[55] Movements of stones, tree branches, leaves and dolls are not actions. When my hand moves because of a strong wind, it is not an action either. Consider the difference between the two questions below:

(q1) Why did the organism's hand move?

(q2) Why did the organism move its hand?

The first question is indifferent to the organism's having a mind or not, or the movement's being willed or not. We can answer it by saying something like "because there was a strong wind, which moved its hand"; or we can as well say "because it wanted to reach for something." However, if we want to answer the second question the same way, we need to say something like: "The organism didn't move its hand. It was the wind that did it." The second question presupposes that the movement was an

[55] For sure, more criteria are needed. When I jump down from the balcony, my jumping is an action, but my falling down is not, even though it is caused by my desire to jump and fall down; placebo effect is a case of mental causation, but the psycho-somatic changes in my body are not actions; my heart beat rises because of a mental state I have, my feeling excited, but it is not an action. As we shall see, these complications are not important for the current discussion. Anyway, these examples already give us reasons to doubt ATS, because mentalistic causal explanation is not limited to actions. It is not even limited to bodily movements. We also causally explain the relation between mental states and how they trigger each other. So ATS, even if it were true, would work only for a limited case of mentalistic causal explanation.

action. In daily life, we generally don't ask questions like the first one. We don't ask "why did that person's body moved in such a way?" Rather we ask why that person did such a thing, looking for a reason or a motivation behind the action, for most of the time people's bodies move because they want them to move. However, this is only a statistical fact. Sometimes, although probably very rarely, our presuppositions fail and we learn that it wasn't the person or the organism that did it. For instance we may wonder why the baby went from this room to that room, only to learn that it didn't really go but was carried by someone else, and we take the question back. The case is not the same for the first question, since there is no such presupposition. If we ask why someone's hand moves and get the answer "because he wills to", we do not take the question back. Rather we get an explanation for what we have asked, and eliminate other possible explanations such as somebody else pulling the person's hand.

Why so many preliminaries about actions? Because it shows us that mental and non-mental states can be used to explain the same thing after all: When we ask "why did the organism's hand move?" we can answer it in the way "because it wanted to reach out for food" or "because somebody tied some tiny ropes to its hand and pulled them." Mental and non-mental states explain the same thing, contrary to the idea of dual explanandum. As we have said, actions are bodily movements caused by mental states, and bodily movements not caused by mental states are bodily movements which are not actions. If externalism is true, it is true that my bodily movements are actions but my swamp twin's bodily movements are not. But this doesn't show that my e-properties bestow powers on me in causing those bodily movements. It doesn't establish that my mental states cause those actions *qua* mental states.

However, let us digress. We might be misunderstanding ATS. What ATS might be claiming is that even though it is the same *thing* that is caused by mental and non-mental properties, mental properties are causally responsible for a certain *property* of bodily movements, which wouldn't be there if the mental property hadn't been there. Let us approach ATS from this angle.

Which property might this be? The only property that actions have but non-actions don't have is the property of *being caused by a state with a mental property*, since that is the only difference between an action and a non-action. It is hard to see in what sense that property is causally explained by invoking the mental property. After all, the bodily movement's having that property *logically depends* on its being caused by a state with a mental property, so invoking it doesn't give us new information and therefore doesn't explain anything. If we know that the movement has the property "being caused by a state with a mental property", we cannot anymore ask the explanatory question of whether it was a state with a mental property that caused it or not, for we presuppose it. Therefore, the "dependence" relation between the two properties is not a causal one, it's a logical one. It is inherent in the way we *define* actions.

Those who are not convinced by the above line of thought may consider this: If ATS were right, the strategy could be used to demonstrate that *every property is causally efficacious for the instantiation of some other property.* Let me explain with an example.

Let’s say that I have a brain state now, and a meteor is passing by in the galaxy 3.000.000 kilometers away from me. And let’s say that if a state has the relational property “existing while a meteor is passing by 3.000.000 kilometers away from it”, we call it an “m-state”, and if a bodily motion is caused by an m-state, we call it an m-motion. Now, one can argue that I can cause m-motions since I have an m-state, but my twin who is only 2.999.999 away from a meteor doesn’t have any m-states and cannot cause m-motions. Two different things are happening in the two cases. The same line of thought taken by ATS could be used to establish that “being 3.000.000 kilometers away from a meteor” is a causally relevant property in explaining what I do.

I do not feel like arguing that being 3.000.000 kilometers away from a meteor is not in any sense causally relevant to my bodily motions, at least not in the above mentioned way. I take it that every common-sensical person can see that the difference between me and my twin, who is not 3.000.000 miles away from a meteor, shouldn’t be covered by any kind of science which is in the business of causally explaining behavior and behavioral capacities. And here is the moral: Nothing changes when we substitute “m-motions” with “actions”, “m-states” with “mental states”, and the property of “being 3.000.000 kilometers away from a meteor” with some e-properties.

I hope the reader is beginning to see more clearly what is wrong with ATS. ATS is offering us some logical or conceptual connections between actions and mental properties, and then tries to sell it as a causal connection.[56]

[56] We should note that relying on actions is rather ironic for externalism. The nature of the mental content behind actions is just the same as the kind of content which has been

What an action is caused by is inherent in the way we define actions. For sure there is a difference between me and my swamp twin: I have bodily motions caused by a state which was caused by H_2O, he doesn't. However, the relation between the property of my brain state of being caused by H_2O and my bodily movement's property of being caused by a state which was caused by H_2O is not a causal one, but a logical, or conceptual one. (See Fodor 1991 for a related point.)

2.4.6 Dretske's Solution: Structuring Causes

We have seen that the action-theoretical strategy doesn't work. It offers us a logical connection between properties as a causal connection. Now we will take a look at another dual explanandum strategy developed over the years by Fred Dretske. (1988; 1995-a, 1995-b) It is in a way similar to ATS, but much more sophisticated and worked out. It is also interesting for our

problematic for externalism: Novel and non-referring content. If mental states that make actions possible have intentional objects, what kind of objects are they? What are they about? Well, they are about forthcoming bodily movements, and since they precede those movements, their content cannot be fixed by causal-informational relations with them. Neither they can be "fetched from the past", since every action is novel, and the content of the mental states behind them are not about "hand movements in general", but a singular, forthcoming hand movement. Teleosemantics will especially have a hard time getting such forward-looking mental states: To get a relevant mental state, it requires a beneficial behavior to make the mental state selected by natural selection either during ontogeny or progeny, so the mental state antecedes the behavior according to the teleosemantic theory. However, in the current case, the mental state precedes the behavior that needs to be beneficial. It is unclear how we can get hold of such mental states within the confines of teleosemantics. Combinatorial externalist theories might find ways to deal with novel, non-referring and creative mental states. However, a combinatorial approach may extinguish the motivations behind externalism: If we can get thoughts and desires about the string theory, unicorns, infinity, and the departmental meeting tomorrow by combination, why can't we get thoughts about water, cows and snakes by combination too? This is not a proper objection in this context, because it only shows this problem makes it harder for externalism, but not impossible, to reach success through ATS. However, I find it a good occasion to send externalists back to the drawing board, and remind the reader about the many other problems the theory faces.

purposes because it was particularly developed in the context of externalism and mental causation, whereas ATS wasn't.[57]

Dretske's reasoning begins in a way similar to ATS. Dretske also claims that while non-mental properties are invoked to explain bodily movements, mental properties are invoked to explain actions, although Dretske prefers using the term "behaving" rather than "acting". According to Dretske, non-mental properties, such as non-relational physical-chemical properties of organisms taken to be non-mental by Dretske, explain why and how their bodily movements happen. However, they don't explain *why* the organism *behaves* the way it does. The why question is supposed to capture *how things got that way,* that is, mental properties explain *how the causal relationship between an internal physical-chemical state and a piece of bodily motion gets fixed.* He calls these kinds of causes "structuring causes", for they explain how things get structured in a certain way.

Dretske illustrates this kind of explanation with a lucid example. (Dretske 1995-b p.112) Suppose that there is a switch that normally rings a bell when closed. We close the switch, only to find out that it turns on a light bulb. We ask "why did the closure of the switch turn the light on?" Here, we are not interested in learning how the switch makes the light turn on, the physical details of the process, we might already know them. What we are interested in is how things got this way. We want to know what *the cause behind the established relationship between the switch and the lighting of the bulb* is.

[57] This means that if I have successfully refuted ATS, I have refuted all its applications. My arguments apply to all, internalists and externalists alike, who would like to adopt ATS as a way to solve the general problem of mental causation.

For Dretske, mental properties are causally relevant in a similar way. We have been through Dretske's externalistic theory of mind before, so let's remember what mental properties according to Dretske are. For Dretske, a state is a mental one if and only if it was caused by a certain thing and if it was naturally selected for its adaptive effects in indicating the thing it was caused by, and indicatory properties explain why a brain state gets naturally selected, therefore explaining how things got a certain way.

For Dretske's strategy to succeed, two things should be established. One is, obviously, to establish that these historical e-properties are really causally relevant in the structuring. But another equally important one is that to establish that structuring causes are the causes we are looking for when we are looking for mentalistic causal explanations. As we will see, none of the two can be established. Moreover, his theory of structuring causes doesn't fit well with his teleosemantic theory about the foundations of the mind.

2.4.7 Are Structuring Causes the Kind of Causes We Are After?

When we laid our problem out in the beginning, we have said that what we are trying to find out is whether e-properties bestow any causal powers on an entity at any given time. To put it another way, we are trying to find out which properties an organism has at any given time endows that organism with certain dispositions, including very complex abilities that enables the organism to perform very complex tasks that helps it to maintain its bodily integrity in a dangerous world. These are the properties we would make use of when we are trying to replicate or simulate the organism, be it in the field of artificial intelligence or a yet-to-come bio-engineering. We also make use of these properties when we are trying to interfere with the mechanism the thing has, to change the course of its behavior. It looks like in cognitive science, we are trying to understand how something works, not how it got that way. We have also said that the debate is not about the causal powers of things in my environment, which they obviously have.

We are also, as a matter of fact, interested in how an organism got that way. Evolutionary psychology and evolutionary biology are examples of such a curiosity. In these fields of explanation, we rely very heavily on the past environments of organisms to understand why they are the way they are. However, these fields are trying to understand how things got that way by means of the causal powers of things in the environment that existed in the past, powers bestowed by properties these things had back then, not historical or relational properties that the organism has *now*. If mental properties as defined by externalism are going to do any causal work, they have to do it now. A mental property I have, like my having a thought about completing my thesis draft in two weeks, is a property that I have *now*. It cannot be used to explain how things got *this* way.

Here, we are beginning to see a strange ambivalence in Dretske's theory. Remember that Dretske reached his theory of structuring causes after concluding that e-properties do not bestow powers on me now in my exhibiting the bodily movements I do exhibit. But if they do not bestow powers on me now, now being a given time, how did they bestow causal powers on me *then* in structuring the connection between my brain states and my behavior, which is just another given time? If our tests with e-properties establish the causal inefficacy of e-properties, they establish it for any given time. My being in the same world with snakes or having a brain state that is caused by snakes do not give me or my brain any causal powers now, nor they gave any back in the day in the structuring of my brain state-bodily movement connections.[58] They do not underlie their dispositions to be selected by natural selection either: The brain states had by Swampman enable him to deal adaptively with his environment even though they do not carry causal information about the environment.

Given this rather simple reasoning, which seems to be true, why Dretske's theory of structuring causes is appealing to him or anyone else? My bet is that it results from a psycholinguistic confusion. It results from the way we talk about certain historical properties which, on the surface, *seem to* attribute causal powers on the historical properties things have now while they actually do not. Let's remember Dretske's example about the light bulb.

When we ask "why does this bulb light up when we turn on this switch?" we can naturally answer "because it is designed that way", which partly explains why things are the way they are, eliminating the other possibility that the structuring happened by an accident, for instance. When we say that something is designed in a certain way, we attribute a historical property to it. We can talk about the histories of objects by attributing them historical properties they have now. But equally, we can talk about the histories of objects by talking about things and events in the past. Instead of saying "because the bulb *is* designed that way", we can also say "because somebody *designed* it that way." Both answers give us the same explanation. It cannot be the case that we are attributing causal powers to

[58] Dretskean relational properties are more specific: Being triggered at the time when snakes are in the environment. What applies synchronic-relational properties apply to them.

different things in the two answers. And as I have argued above, it is the properties of the designer who designed the circuit that was causally relevant for how things got this way. The historical properties of the switch that it has now are causally silent.

What enables us to speak in both ways is the fact that historical properties things have at a time *logically entail* certain properties or events existed in the past. If the circuit has the property of "being designed" at T1, it entails that there was an event of somebody designing it before T1. If someone designed a circuit at T1, then after T1 the circuit will have the property of "being designed", stuck to it forever.[59] What I offer as an explanation of the initial appeal of Dretske's theory is this: We sometimes explain things by talking about properties something has now, which might create the cognitive illusion that it is the current property that does the explanation, even though our acceptation of the explanation is due to the properties that existed in the past entailed by the current properties which are talked about.

2.4.8 The Explanatory Strategy

Refuting Dretske's theory of structuring causes helped us to get clearer about causal explanation. We have seen before that sciences that try to explain *how* things work are interested in some properties that are not historical and not broad relational properties. Now we have also seen that sciences that try to explain *why* things are in a certain way are *interested in*

[59] But note that the entailment is not the other way around. If there is an event of someone designing a switch at T1, this doesn't imply that there exists a switch now with a historical property of being designed, for the switch may have ceased to exist.

history, but not historical properties. They are interested in properties that things had before, but not their historical properties that they had then or now. I also tried to offer a psycholinguistic explanation of why the theory of structuring causes might have a prima facie appeal to anyone even though it is clearly mistaken.

Now we are done with the kinds of objections which state that me, my swamp twin and my dry twin do not have the same causal powers. The next objection we will consider will be of a more general sort, leveled against the current methodology, which is the last resort of externalism to establish that e-properties are causally efficacious.

One line of thought against the methodology that is adopted in this thesis is that when we are trying to see whether something can figure in a causal explanation, our starting point should be successful explanations in science and everyday life and *not some a priori metaphysical principles*. This idea has been most famously urged by Burge (1986, 1995) and Baker (1993). It hasn't always been put forward and discussed in the context of externalism, but generally as a response to the general problems related to mental causation. However, Burge and Baker has also deployed it to save externalism, and it's a significant idea that anyone can potentially deploy against the argument developed in this thesis, so it needs to be properly dealt with.

Here is an often quoted passage from Baker:

> My suggestion is to take as our philosophical starting-point, not a metaphysical doctrine about the nature of causation or of reality, but a range of explanations that have been found worthy of acceptance.

> Construing explanations as answers to 'why' questions, with perhaps some constraints on what count as an adequate answer, my proposal is to begin with explanations that earn their keep, rather than with the metaphysics, which seems to me a freeloader that just interferes with real work. (1993, pp. 92-93)

The pursuers of the explanatory strategy might be onto something. If a couple of philosophical arguments try to show us that we are wrong in pursuing some explanatory practice which proved useful for some hundreds of years, the first thing we should do is not to give up the practice, but to check what might be wrong with the philosophical arguments. Not only because the well-worn practice has a high chance of turning out to be on the right track, but philosophical arguments are notorious for generally being on the wrong track.[60] The strategy also accords well with a conservative principle in philosophy of science, which I find plausible: When there is some ill-fitting data, a scientific theory should be able to resist change before first reconsidering and reevaluating the data, including philosophical data.

As the reader can guess, I will defend myself against the explanatory strategy. But let me begin with some general worries before moving on to my actual objections, which will also help us in understanding in which way the explanatory strategy is supposed to work.

[60] At this point, let me just note that there seems to be a double standard on the side of Burge, for the very explanatory arguments he has provided against internalism could have been provided against externalism when it was emerging. It is common to hear that externalism, when it first hit the scene, was very unusual for philosophical and scientific tradition. Like Burge and Baker, one could defend the internalist tradition on explanatory grounds and try to see what might be wrong with those very few thought experiments that won externalism its name, like Putnam's Twin Earth and Burge's arthritis. But things didn't happen this way, which is a curious historical point that I can't pursue further here.

It seems like a plausible idea that we should attend to real, readily working causal explanations when thinking about matters causal. However, saying that we shouldn't heed philosophy doesn't solve the problem, for we have already confronted the philosophical questions and acquired our worries. Let's assume for a moment that successful causal explanations in science and in everyday life are really externalistic. On the other hand, we have plausible arguments to the effect that e-properties are irrelevant for causal explanation. The argument I have given in this thesis might be false, but I take it that it is *plausible*, and definitely not silly, and any person who claims that e-properties are causally efficacious should be faced with a puzzle upon encountering them, which should leave them in some "cognitive dissonance" until the puzzle is solved. One should ask oneself: "Ok, our everyday practice seems to tell us that e-properties are doing some causal work. But *how* can they do it, given some plausible arguments to the contrary?"[61]

So there seems to be two paths to take in such a situation. One can provide counterarguments for the a priori metaphysical thesis, but this would just be what everyone else has been doing, and no mentioning of the pragmatic or epistemological matters is needed. One can also choose to live with the puzzle, but this is rather undesirable, and the third route, just ignoring the problem, is not a good intellectual response.[62] But let's leave these worries

[61] This worry regarding the explanatory strategy has been put forward in Kim 1998 (pp. 60-7), and the usage of the term "cognitive dissonance" in this context is also due to him. (Kim 1989)

[62] One could also have a worry like the following: It is always possible that there might be something wrong with our explanatory practice. It is a natural and good-sounding idea that we should follow our most successful explanatory practice, but this doesn't mean that when we face some problems, we shouldn't examine and revise them. Old exemplary

aside. Let’s say that someone chooses to ignore the problem or to live with the dilemma.

Note that the explanatory strategy does not tell us where the philosophical theory goes wrong. It seems to suggest that if the best causal-explanatory practice tells us that e-properties are causally efficacious and if some philosophical theory says that they are not, then there *must be something wrong* about the philosophical argument, although we don’t know where, and the current causal-explanatory framework *shouldn’t* be replaced. If this analysis of the explanatory argument is correct, it is better to view it as a normative argument about which way we *should* go when faced with such a dilemma.

Having said all these, let us reconstruct the explanatory argument this way:

> (ES1) All ordinary and scientific successful causal-explanatory practices mention e-properties as causally efficacious.
>
> (ES2) If certain properties are mentioned as causes in a successful causal-explanatory practice, their causal inefficacy can only be shown by some other causal-explanatory practice, not a priori considerations about causation or explanation.

frameworks are replaced with better ones. It is always possible to find out that our current framework is inadequate, its posits non-existent, and its assumptions inconsistent. Blindly relying upon the current practice looks dogmatic, and it is a very easy way to go: It could be used to defend *any* theory, but we know that some theories are eventually replaced. However, friends of the explanatory strategy can block this worry, by saying that successful-looking theories can surely be replaced, but they are to be replaced when *other* successfully proven explanatory frameworks hit the scene, not when someone comes up with some a priori philosophical concerns about causation. I will consider this line of thought later on.

Therefore,

(ES3) The establishment of the causal inefficacy of e-properties by the comparative method, which is based on a priori considerations and not on our best causal-explanatory practice, cannot establish that e-properties are causally inefficacious.

We could refute the argument simply by showing that (ES1) is not sound. But before that, let me deal with some other problems inherent to the argument.

2.4.9 Can a Causal-Explanatory Practice Be Displaced By Something Other Than Another Causal-Explanatory Practice?

The explanatory strategy tells us that we cannot rely on a priori principles about causation in establishing the causal inefficacy of e-properties. But what makes some considerations a priori, rather than a posteriori? We can say that the causal efficacy of a property can be established in an a posteriori way only through *experimentation*.

This leaves us with two questions, the first one being the following: Are the mentioning of e-properties, in our successful causal explanatory practices, if there are any, free of any non-experimental background? Surely it is not, in two ways. Firstly, every causal-explanatory practice is grounded in a non-experimental understanding of what causation or causal explanation is, and that understanding can well be confused. Secondly, externalism itself is a philosophical theory established a priori through philosophical method.

Surely we cannot find out through a priori methods things like whether this property of the brain or that one is causally relevant in my going to the kitchen. We cannot find out, from the armchair, whether this gene or that gene plays a causal role the development of my eyes. Philosophy is not in this league. The a priori contribution of philosophy is to provide an understanding, a conceptual analysis of causation. In this thesis, I have been relying on a few simple and pretty much uncontroversial ideas about causation, such as the idea that if something is causally efficacious we can in principle intervene with it to change the course of events, something that is anyway behind our "best current explanatory practice". These ideas guide us in *making sense* of causation, which in turn lets us see whether e-properties can be causally efficacious or not, or even their causal efficacy can be *coherently imaginable* in the first place. It is true that these ideas are a priori: Without them it is hard to see how we can make sense of causation at all and go on to do our a posteriori business. Their falsity cannot be established by empirical methods, for such establishment wouldn't make sense. One cannot say that she has found out, through experimentation, that there are causally efficacious properties that do *not* make a difference, for when there is no difference around, there is no experiment to speak of.[63]

If our Baker-Burgean objector needs to show us that the causal efficacy of e-properties is not established a priori she needs to show us a case of experimentation where e-properties make a difference. However, that's what I have shown to be impossible. The Burgean objector can of course try to show us that e-properties are causally efficacious by trying some of

[63] To be fair to Burge, the "a priori metaphysical principles" he focuses on in his 1993 and 1995 are the principles of physicalism and the causal closure of the physical.

the routes we have been through, such as the action theoretical strategy or Dretske’s theory of structuring causes. But then she would end up in a priori waters again, for these strategies just rely on a priori methods when trying to find out what is causally relevant for what. For instance, Burge himself, claims that in a mentalistic causal explanation we are interested in what causal information is carried by organisms and their bodily movements, and that is what e-properties make a difference in. (Burge 1986) I, for one, am not interested in what causal information is carried by the bodily movements of organisms when I am looking for a mentalistic explanation. Burge can criticize me about this, but this only shows that we are battling in the a priori waters.

So it seems that for the Burgean strategy to be forceful, its first premise should be modified this way, removing the mentioning of the a priori:

> (ES1*) All ordinary and scientific causal-explanatory practices mention e-properties as causally efficacious, and the causal efficacy of these properties is established by experimentation.

However, that is what exactly is at stake. For saying that a property’s causal efficacy to be established by experimentation is just saying that it is causally efficacious, for the causal efficacy of something can only be seen through differences it makes. If we had this premise in an argument and conclude that e-properties are causally efficacious, the argument would presuppose its conclusion, and wouldn’t be able to get off the ground.

At this point, our Baker-Burgean objector can try to simplify her argument, to not to get into the above mentioned problems:

> (ES1**) Ordinary and scientific successful causal-explanatory practices mention e-properties as causally efficacious.
>
> (ES2**) If certain properties are mentioned as causally efficacious in a successful causal-explanatory practice, they are causally efficacious, no matter what else seems to point to the contrary.
>
> Therefore,
>
> (ES3**) E-properties are causally efficacious.

(ES2**) is a very strong premise. It completely closes all the doors for a philosophical criticism of a successful theory. The big problem with the premise is that it doesn't tell us what to do if more than one equally successful framework based on the same empirical evidence claim different things. Think about the dozen or so interpretations of quantum phenomena that seem to fit the data equally. Most of these interpretations are necessarily false. Similarly, the idea that e-properties are causally efficacious can be equally false even if there is a successful theory where e-properties are mentioned as causes.

More generally, it is unclear why anyone should hold (ES2**) to be true, given how strong it is. It is possible to think of many ways how can parts of a theory be false even though it is successful, how can e-properties figure as causes in a successful theory even when they are not causally efficacious: In a successful theory e-properties can be redundant; they may be reliably co-varying with the properties that are really causally efficacious; they may be waiting to be replaced by something that does a better job; or there might be a conceptual confusions on the side of the theorist.

I hope all these give us enough reasons to doubt the explanatory strategy. For those who are still not in doubt, I will also take a critical look at the first premise:

> (ES1) All ordinary and scientific successful causal-explanatory practices mention e-properties as causally efficacious.

I will try to show that this is false. The word “all” in the premise is important. What to do if there are some successful theories that give us everything we want from a mentalistic causal explanation without mentioning e-properties? There are, indeed, successful theories that do not mention e-properties, and given that we have both kinds of theories, the explanatory strategy does not work. But would it work if there weren’t any internalistic theories around?

2.4.10 Is Everyday and Scientific Causal Explanation Really Externalistic and Does it Matter?

To vindicate the causal efficacy of e-properties, one needs to point to certain successful theories in science and practices from everyday life where mental externalism is assumed. Hopefully, Burge didn’t leave his strategy ungrounded and tried to find some good looking theories in science where, allegedly, mental states are individuated externalistically.

Burge took up David Marr’s theory of vision. A good deal of debate has been produced between Burge and Segal, and some other authors, about whether the individuation of mental content in Marr’s theory was externalistic or not. (Burge 1986, Segal 1989, Egan 1991) I do not want to join this specific debate, nor do I want to review it. I believe that Marr’s theory was too new at the time of the debate to be regarded as a successful theory, and moreover, it was a paradigm work in a newly emerging theoretical background in mind sciences, the methodological and conceptual foundations of which have been hotly debated and continues to be debated to this day. Also, Marr died before seeing this debate, even before the publication of his influential book in 1982, so he didn’t have a chance to respond to the debate between the philosophers. It might be the case that Marr’s theory is totally indifferent to the truth or falsity of externalism. If Burge can interpret the theory in an externalistic way and Segal in an internalistic way, then it seems that there is a lot of interpretive space open, and the theory is silent about the foundations of mental states.

It can be questioned whether trying to decide on the issue looking at very specific theories by certain scientists is a good way. Now, what can we establish if, as a response to Burge, I take up Chomsky’s theory of generative grammar, who is explicitly internalist about his assumptions? (Chomsky 2000) The debate would naturally end in a stalemate.

Let us grant that Marr’s theory is a successful one, and let us also grant that Marr’s theory is an externalistic one and Marr was even explicit about this. However, as I have said, the success of a theory does not automatically vindicate its theoretical commitments, and it seems that one can save many empirical generalizations acquired by the theory while correcting some of its assumptions. A philosopher has the right to investigate these issues and warn the theorists, pointing to assumptions that are redundant, that are unjustified by the theory itself, that contradict with the implications of the theory, and which rest on conceptual confusion.

Let me give an example: Cognitive neuroscience came up with many empirical findings to this day, such as correlations between certain cognitive tasks and neural states or the chemical processes subsuming memory, and these findings seem to aid us in explanation, prediction and control, which makes the theory a successful one. Also, very same neuroscientists today say many things that are taken by philosophers as problematic: They say that the mind is the brain; that memories are entities stored in the brain; that the brain processes information; that persons are their brains; that there are images in the brain that we perceive, rotate and inspect; etc. Bennett and Hacker (2003) attack all these claims on conceptual grounds, arguing that these claims doesn't make sense, while also pointing to the fact that many successes of modern neuroscience can remain intact, and even more properly acknowledged, without these assumptions. I am not mentioning this to support Bennett and Hacker's specific claims, but just to point to the possibility that if an empirically successful theory is even explicitly externalistic, its success can be kept untouched while doing away with externalism. So, a defender of the explanatory strategy should not only point to a good theory's being externalistic, she also needs to show that externalism is *necessary* for the theory to work.[64]

[64] Also note that many cognitive neuroscientists explicitly say that the mind depends on just the current state of a brain. So, it seems like one could try to save internalism by just the same strategy deployed by Burge and Baker.

So, it looks like the explanatory argument is a non-starter. However, some readers might be still wondering whether the best theories we have today in science that explains behavior, or folk theories in that case, are externalistic. I will not cover this issue in detail, but let me give a somewhat anecdotal short list.

As we have seen, Chomsky's linguistic theory is an explicitly internalistic one. It is not committed to the existence of words or noun-phrases in the mind-independent world; many cognitive neuroscience textbooks will tell you that it is theoretically possible that you can be a brain-in-a-vat; subjectivism about sensory qualities is a near-consensus in psychophysics; those who study development of conceptual categories don't care about whether those concepts refer. Neither in mainstream psychology or psychotherapy a practitioner feels a need to search for the causal-informational history of her patient to tell her how she should overcome her jealously, not even with the fact that the patient's husband is really cheating on her.

Although not a causal-explanatory one, lexicography is an internalistic science. Unlike an encyclopedia, dictionaries don't care about mind-independent world. A dictionary doesn't tell you whether the word "unicorn" or "water" refers or not. A dictionary, unlike an encyclopedia, is interested in "mere meanings", not "the world." A book that documents various religious beliefs and practices around the world doesn't tell you whether any of those religions are right or wrong in postulating the supernatural entities they postulate, still they make you understand what people believe and why they behave the way they do. Similarly in everyday life, we seem to understand each other's and our own beliefs and desires without caring to know about the causal or etiological roots of these mental states or if their content corresponds to anything in the world, and predict each other's behavior according to those attributed mental states.

I think this quite casual list is enough for the reader to make her doubt whether well-working practices are externalistic, and that's all I aim at. It should be noted that the internalistic style of mental state attribution is not subjectivistic, rather, it is an instance of methodological solipsism: It is possible to understand one's mental states while being totally *indifferent* to the question of what exists or existed in the mind independent world. For some this might sound quite intuitive. After all, psychology is interested in what is in the mind, not what is in the world. Twin Earth externalism wants us to mix up psychology and linguistics with the chemistry of water. But this division of labor in the sciences is not so intuitive to externalists, and that's why we should move on.

2.4.11 Can E-Properties Be Causally Relevant without Being Causally Efficacious?

If my arguments in the previous sections are sound and valid, and if there is no interesting objection that we have missed, then we can safely conclude that e-properties are not causally efficacious.

In the beginning we have noted that causal inefficacy does not automatically lead to causal irrelevance. There is a possibility that e-properties can be causally relevant in an *epistemic* or *pragmatic* way. This can be in two ways: If knowing e-properties gets us closer to the properties that are really causally efficacious, or if knowing them allows us to make predictions. I take it that these are possible only if e-properties systematically co-vary with those properties that are causally efficacious.

If two properties X and Y lawfully co-vary with each other, and if X is causally efficacious in bringing about Z while Y is not, then it is still possible to predict the occurrence of X if we have knowledge of Y and don't have the knowledge of X. If every creature with a kidney is a creature with a heart, and if hearts produce heartbeats, then we can predict that some beating sounds will come out of the chest of a creature with a kidney, even if we don't know anything at all about hearts, and even though having a kidney plays no role in the production of the sound.

Before trying to understand whether this really is the case or not for e-properties, let us note a few things. Even if e-properties are causally relevant without being causally efficacious, this is still a huge blow for externalism's place in an explanatory science, for the possession of these properties would not explain behavior, and therefore they would not let us understand how and why organisms behave the way they do and what enables them to do the things they do. Even if e-properties are causally relevant only in a pragmatic or epistemic way, it does not secure their place in a complete explanation. Mentioning them would not satisfy our epistemic hunger for causal explanation.

One theoretical proposal that aims to find the causal relevance of e-properties this way is the *program explanation* theory developed by Frank Jackson and Philip Pettit, through a series of papers written in the late eighties and early nineties. (Jackson & Pettit 1988, 1990) The theory is not only developed with e-properties in mind, but it is developed as a more general framework to understand the place of certain properties in explanation.

Jackson and Pettit, at least as the authors of the above mentioned papers at that time, are two-factor theorists regarding mental content. They hold that mental states have both broad and narrow content, broad content working in an externalistic way and narrow content not. Moreover, they hold that *both* narrow and broad mental properties are causally inefficacious, thinking that it is some non-mental properties doing the causal work in behavior, such as certain properties that can figure in microphysics. However, they hold that mental properties are causally relevant in the sense that their existence *guarantees* the existence of the causally efficacious non-mental properties. For this relation they use the term "programming", a term that I find rather misleading. By pointing to a mental property, we *disqualify* certain other explanations, explanations that would rely on properties that do *not* co-vary with those mental properties.[65]

Jackson and Pettit propose that wide content gets us closer to narrow content, which gets us closer to the properties that are causally efficacious. For their proposal to work, it should really be the case that e-properties should get us to closer either to narrow content, or the effects that follow from a given narrow content. However, as I will try to show, for a mental state analyzed in an externalistic way, there is no way to get to the narrow content from the broad content.

[65] See also Yablo 1992 for a similar theory, who holds that the relation between mental properties and causally efficacious properties are like the relation holding between determinates and determinables.

Take for instance, “a belief about H_2O” or “the concept of H_2O”, understood in an externalistic way. According to externalism, it is correct to say that one can have a belief about H_2O even though she knows nothing about H and O. Moreover, broad content does not depend on *what* one believes about H_2O. Someone who knows chemistry and who has the belief that water is potable can have the very same broad content as a person who doesn’t know chemistry and who believes that water is poisonous. As we have seen before, it is what these people believe about water that makes us predict their behavior, not what it refers to. For all we know, someone who has a brain state caused by water can believe that water is a type of stone found on Uranus.

It is true that having broad content implies having *some narrow content or other*, even though it doesn’t tell us *which* narrow content it is. By citing broad content, we eliminate certain other possibilities. For example, by saying that someone reached out for H_2O because he had an internal state caused by H_2O, we eliminate the possibility that her movements were caused by non-internal states. But the elimination of non-internal states is possible only because we mention that there is some or other internal state. The property “caused by H_2O” does no eliminative work here. Saying that she had an internal state caused by XYZ would do the same work; it would eliminate certain external things, not because it is the XYZ doing the causing, but because we mention that it was the internal state and not something else. There is no reason why we shouldn’t chop off the e-properties and just mention that there is some internal state or other, if it is necessarily true that an e-property co-varies with some or other internal state, and if it doesn’t tell us which state it is. Talking about e-properties is just redundant.

In a nutshell, knowing the property “being caused by X” does not give us any predictive power if we have no idea *what is the thing it had caused*. On the other hand, if “being caused by X” necessarily implies some very general category like some internal state or other, then in each case we would talk about the e-property, we can just talk about what is implied by it to avoid redundancy.

2.4.12 Fodor’s Proposal

A proposal have been put forward by Fodor in his 1994, a proposal that tries to show us why we shouldn’t be worried about the causal problem for externalism. We should first note that Fodor generally frames his theory not in terms of causal powers, but of *law-like generalizations*, and it is why it is included in this section. For him, e-properties will secure their place in science if it is possible to make law-like generalizations via them, and if there aren’t any cases in the world that will force us to look for something else that broad content cannot cover. Before introducing the proposal, it will be fruitful to take a look at Fodor’s earlier views about the matter, which will make us better understand his final position.

In the eighties, Fodor has been a defender of non-externalistic narrow content along with broad content on causal-explanatory grounds, for he thought that it is only narrow content that can be causally efficacious or allowing for law-like generalizations. One reason that made him think this way was twin cases and Frege cases. In his 1994, he proposes that we may not need narrow content and we can do law-like generalizations based on broad content if our world is such that twin cases and Frege cases do not exist.

Fodor takes the Twin Earth example. He admits that if there were cases like the Twin Earth case, then e-properties would fail to capture certain law-like relations, for we would need to introduce some narrow content that will cover both Earthians and Twin Earthians. But he says, for all we know, there is no XYZ in the universe, or similar cases that we find in the thought experiments (and as we will see, in works of literature) that make us worry about the causal status of externalism.

Unfortunately for Fodor's proposal, Twin Earth cases, Frege cases, Dry Earth cases and Swampman cases *do* exist in our world, as I will try to show. Let's begin with Swampman cases.

2.4.13 Twin Earth, Dry Earth, Frege and Swampman Cases in Real Life

Remember that Swampman is a creature who comes into existence by chance. Surely, we do not have such full creatures coming into existence by chance. Nevertheless, we have swamp body parts and states, such as brain parts and states, which come into existence by chance, that lack e-properties relevant for externalism. Think of first mutants in the biological world. A first mutant is a creature that gets a genetic mutation for the first time, a mutation which in turn causes changes in its body parts during the course of its development. If mutations happen by chance, then cases of first mutants are miniature swamp cases.

Let's imagine that through a mutation, a creature is endowed with the ability to behave adaptively towards snakes with the brand new brain state it has. The brain state, when the creature first has it, lacks the property "being caused by snakes" and "being selected for its adaptive effects in snake-oriented behavior." According to versions of externalism, this creature does not yet have snakes in mind. But necessarily, the creature has the ability to behave appropriately towards snakes *before* the brain state gets naturally selected for its adaptive effects in snake-oriented behavior. The two brain states of the creature, the one before it gets selected and after it gets selected and acquires the e-property, is just like Swampman and Man, respectively. E-properties fail to capture the swamp states, forcing us to look for a common property shared by swamp cases and non-swamp cases.

Dry Earth cases are everywhere. The imaginary case of Dry Earth is no different than mundane versions of false belief and non-referring concepts. For people ending up in the middle of the Atlantic Ocean to look for Atlantis, e-properties defined as their relations to Atlantis cannot be used to form law-like generalizations to capture their behavior, since there is no Atlantis to stand in a relation to. This forces us to look for something common between people who end up in the middle of the Atlantic Ocean and who end up in Atlanta.

What about Twin Earth? Can we find examples of people who cannot differentiate between two things although the two things are different? Obviously we can, and they are ubiquitous. Jade and jadeite is the paradigm case we come across in the literature, but we don't need to go that exotic. We can think about any everyday objects that cannot be distinguished. Even for water, which was the example used in Twin Earth thought experiment, we have H_2O and varieties of heavy water: D_2O, ^2H_2O, HDO and $^1H^2HO$. The existence of these cases force us to look for something common between people who only came across Jade in their lives and who came across Jadeite, and people who only came across Obama and people who came across his stunt.

Frege Cases, which are the mirror-images of Twin Earth cases, are another species of mundane phenomena, where people believe that there are two things even when there is one. The prototypical Frege case of the Evening Star and the Morning Star is a real life case. I believe the reader can find more examples. These examples force us to look for something different in people who have the same e-properties (such as “having a brain state caused by Venus”), but who are different Frege-wise.

Fodor himself mentions the mythical case of Oedipus, a case that creates problems for externalism. Oedipus’ case is a Frege case. There is only one woman around, but Oedipus thinks that there are two: His mother, and the women he is in love with. He doesn’t know that the woman he is in love with is his mother. The case presents a problem for externalism because it seems that we need to attribute two different mental states to Oedipus, thoughts about his mother and thoughts about the woman she’s in love with, to properly understand and predict his behavior. On the other hand, a referential externalism attributes to him only one kind of thought. However, if we say that his thoughts refer to his mother and therefore in both cases he has thoughts about his mother, we have to say that he wants to make love with his mother. But this is rather unintuitive, and Oedipus would definitely answer “no” if he was asked “do you want to make love with your mother?”

However, Fodor tells us that externalism shouldn’t be worried about Oedipus like cases, for they are mythical. He thinks that the reason why Oedipus makes an interesting story is that it is highly unlikely. In the everyday mechanisms of the normal world, there is something which guarantees that such Frege cases don’t arise, that people generally know who their mothers are and recognize their mothers as their mothers when they see them. So, one kind of content, of an externalistic sort, will always work. And we will be able to formulate the *ceteris paribus* law-like generalization that if someone believes that a person is his mother, he will avoid making love with her.

Fodor’s conclusion is too quick. Cases like Oedipus is pretty rare for sure, but other ones, such as the Venus case, are not. Even for those that are rare, they are definitely not impossible. There is no reason to think that on this very Earth, there haven’t been people who wanted to make love with a woman without knowing that she is their mother, given the existence of many cases of incest, adoption, separation, or even encounters with unrecognizable youthful pictures of one’s mother. Oedipus cases are rare, but this is a contingent fact. They are rare because of many reasons: Children generally spend a lot of time with their parents and recognize them when they grow up; it is unlikely for the separated children to encounter their mothers again in this big world; it is unlikely that they will fall in love with that particular women among many people, who are, anyway, much older than themselves and therefore not sexually attractive. This is why the story of Oedipus is fascinating. But there is no nomological fact about the world that makes the story impossible. If the whole world was like the early Soviet Union where the family institution was being abolished, if the world were smaller, if there were fewer women, if more people had an interest in elderly women, then Oedipus cases would be more likely. And nevertheless, there are Venus-like Frege cases, which are quite likely.

The theoretical background of a scientific paradigm definitely shouldn’t be based on such contingencies that are here today and tomorrow not. For his argument to work, Fodor needs nomological impossibilities, not contingent facts. But there are no such impossibilities. Therefore, Fodor’s argument to the effect that we have no motivation in the real world to do away with narrow content is ungrounded.

CHAPTER 3

CONCLUSION

In this thesis, we asked the question: “Can externalism in philosophy of mind be the theoretical backbone of a science that wants to mention mental properties as the causes of certain effects?” The answer turned out to be a qualified and well argued “no”. How did it “turn out” that way? Did we find the answer in the recent findings of the natural sciences? No. Our question was not the kind of question that could be solved that way. It was a conceptual one.

Every explanatory scientific framework, when it goes on to do its empirical business, already embodies an understanding of what explanation is, what causation is, what relational and non-relational properties are and how they work. The questions regarding what these things are cannot be found out by looking at the findings of those sciences. These concepts are very basic concepts at the heart of our understanding of the world.

When we look at any scientific framework, such as the causal-information theoretical framework of the mainstream cognitive science, it might seem to us that some historical, relational properties are doing some causal explanatory work, while this might not be something more than a conceptual carelessness on the side of the theorist, something that itself contradicts with some other assumptions of hers.

A cognitive scientist may claim that having a mental property depends on things that happened in the history of the organism, and having mental properties are instrumental for a complex organism to survive. But when presented with examples, she will also find himself claiming that a Swampman without a history, a biological mutant with a brand new brain state, a person with neural enhancements and a robot that came fresh out of the factory will all have the same ability in survival as their counterparts that have the right kind of history. A cognitive scientist may claim that mental properties are relations to things that exist in one's world, and these mental states explain what people do, but then they will find themselves claiming that the existence or the non-existence of unicorns has nothing to do with one's powers in drawing picture of a unicorn.

What shall do when we are faced with such conflicts? One answer that can be given to this question is that, the beliefs that shall stay are those beliefs that are more deeply entrenched, those beliefs that which if we give up on, a large part of our understanding of and the manipulation of the world will crumble. The idea that two entities with the same intrinsic physical chemical properties should have the same causal powers, or the idea that two people who believe that there is a potable liquid which flows from taps are disposed to go the basin when they think they need some of that liquid, is definitely more deeply entrenched and central to our world dealings than the recently introduced philosophical idea that mental properties are relations between one's mental organs and to things that exist outside of those organs, an idea, however being popular, rests on a trio of questionable thought experiments.

A philosophical tradition that began with Quine has been telling us that no degree of deeply-entrenched-ness is a proof of immunity to revision. This might be true or not, but we see nothing in the world to make that revision and enter that alien conceptual landscape externalism wants to push us into. It is not what something is caused by, but what something is, that tells us what that thing can do. Similarly, it is not what else exists in one's world that tells us what powers one has and how she has those powers; after all, everything else in the world has that property of sharing a slice of reality with that mind-external entity. We will have no need to believe that the instantiation of properties offered by externalists have any causal effects on the world, not until the day that we find objects that cannot hold ashes because they were not designed to hold ashes, people who cannot draw pictures of unicorns because they never adaptively acted on unicorns, organisms that cannot reach out to water because they don't have any bodily states caused by water, people who stop going towards the way of Danube River right at the time she dries out.

As we have seen, there are attempts to save the externalist theory against this causal challenge, trying to open up a space for e-properties in the causal chain of the world. But when we examine them, we find nothing but conceptual sin. To count them, seven sins:

(1) The idea that e-properties overdetermine behavior sins against the fact that e-properties are not sufficient for the production of behavior, and even if it somehow cleanses itself of this sin, it is

suspicious for being an ad-hoc response that is pathologically closed to refutation.

(2) The idea that one doesn't have the ability to drink H_2O if one doesn't have H_2O in one's planet sins against the fact that causal powers are tested in the same contexts. The similar idea that one doesn't have the power to drink H_2O if there is no H_2O in the total reality sins against the fact that causal powers can be tested counterfactually, and dispositions can be targeted at things that do not currently exist. Also, both ideas are in conceptually troubled waters for claiming that my being able to drink H_2O is a power of mine even though something outside of me, the sample of H_2O in the bottle in my hands, is something outside of me, and for trying to this while not collapsing into the extended mind theory, which is different than content externalism.

(3) The action theoretical strategy sins against the fact that the relation between mental properties and the property of being an action is not a causal one, but a conceptual one, while trying to sell that connection as a causal one.

(4) Dretske's theory of structuring causes sins against the fact that historical causal explanation, such as in evolutionary biology, is different than causal explanation which tells us which properties of an entity are responsible for its causal powers and how, such as in mechanistic biology. Dretske seems to forget the fact in the debate about the place of e-properties in cognitive science is about the second type of explanation. Moreover, it is not obvious if Dretske's e-properties can have any place in the first one either.

(5) The explanatory strategy exemplified by Burge and Baker sins against philosophy itself by trying to make our theories untouchable by philosophy, forgetting the fact that the mere mentioning of e-properties as causes in our best explanatory scientific and common-sense frameworks does not justify that e-properties for these theories might suffer from internal inconsistencies, confusions and linguistic carelessness which can be cured by nothing but philosophical method. Even if the strategy is supposed to work, it is not obvious which side it will work for, as many internalistic assumptions loom large in our best theories.

(6) Jackson and Pettit's theory of program explanation tries to find causal relevance for e-properties without causal efficacy, and sins against the fact that e-properties do not systematically co-vary with other properties that do the causal work and therefore do not get us closer to the causally efficacious properties.

(7) Fodor commits the most evil of sins when he blinds himself in an ad-hoc fashion to the real world counterparts of Frege cases, Twin Earth cases, Dry Earth cases and Swampman cases which are problematic for his attempt to find a nomological relationship between e-properties and behavior, and for doing this as a well-heeded captain of the ship of theoretical cognitive science, his sin is unforgivable.

If it is quite simple to find out that externalism is incompatible with causal explanation, how did externalism became so popular in the philosophy of cognitive science? That is something to be explained by history and sociology of science. At some point in history there appeared the Twin Earth thought experiment, and the climate was ready for it to spread like the common cold before being examined. Twin Earth got there first to pump our externalistic intuitions, before Swampman or Dry Earth which pump our internalistic ones. At some point, we were tricked into believing that mental states work like linguistic representations. In more recent times, people came to believe that "externalizing" the mind would solve a lot of philosophical problems, and help us get rid of the unwanted legacy of Descartes.

But Descartes was right, at least about one thing. After all these times, it is still possible to imagine that we can be tricked by evil demons, brains-in-vats or characters in Matrix - living in a solipsistic world. Bringing up this conceptual possibility is of course not aimed at showing us that we actually are living in a solipsistic world. It is aimed at showing us that what one has in one's mind is independent of what exists outside of it. Water is in the world, and it is also in the mind. If one day all water ceases to exist, or if it turns out that it never existed, then we will say that it is *merely* in the mind. Chemistry should deal with water, or H_2O and XYZ, what is in the world. Psychology should deal with water-thoughts, what is in the mind. It should be, as it is called in the trade, methodologically solipsistic.

My behavior and the behavior of my soul brother in Twin Earth are both explained by this water-thought. They have the very same content, they involve the same concept: WATER. To learn about the properties of water, we go and take a look at water. Chemists on Earth and Twin Earth will find different things. But when we want to find out what WATER is, what people have in their minds, we simply go and ask people, and on both planets we find the same thing in people's minds. Externalism blurs this distinction, and introduces an alien norm which tells us that we should postpone psychology to the days of completed chemistry to correctly understand and talk about the thoughts of our own and our fellow Twin Earthlings.

BIBLIOGRAPHY

ADAMS, F. & AIZAWA, K. (2008) *The Bounds of Cognition.* New York: Blackwell.

BAKER, L.R. (1993) Metaphysics and mental causation. In Heil and Mele (1993)

BENNETT M. R. & HACKER P. M. S. (2003) *Philosophical Foundations of Neuroscience.* Oxford: Blackwell.

BERMUDEZ, J.L. (2003) *Thinking without Words.* Oxford University Press.

BILGRAMI, A. (1992) Belief and Meaning. Cambridge, Mass.: Blackwell.

BOGHOSSIAN, P. (1997) What the externalist can know a priori. *Proceedings of the Aristotelian Society*, vol. 97:161-175.

BRENTANO, F. (1874/1973) Psychology from an Empirical Standpoint. London: Routledge.

BURGE, T. (1979) Individualism and the mental. *Midwest Studies in Philosophy* 4. 73:121.

BURGE, T. (1982) Other bodies. In A. Woodfield (ed.) *Thought and Object: Essays on Intentionality*, pp. 92.129. Oxford: Clarendon Press. Reprinted in Pessin and Goldberg (1996)

BURGE, T. (1993) Mind-Body causation and explanatory practice. in Heil and Mele (1993)

BURGE, T. (1986) Individualism and psychology. *Philosophical Review* 95: 3-45. Reprinted in Macdonald and Macdonald (1995)

BYRNE, A AND LOGUE, H. (eds.) (2009) *Disjunctivism: Contemporary Readings*. MIT Press.

CHOMSKY, N. (2000) Internalist investigations. in N. Chomsky, *New Horizons in the Study of Language and Mind.* MIT Press.

CLARK, A. (2008) *Supersizing the Mind: Embodiment, Action and Cognitive Extension.* Oxford University Press.

CLARK, A. & CHALMERS, D. (1997) The extended mind. *Analysis* 65: 1-11.

DAVIDSON, D. (1987) Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* 60: 441-58.

DESCARTES, R. (1642/1996) *Meditations on First Philosophy, with Selections from the Objections and Replies*, trans and ed. J. Cottingham. Cambridge University Press.

DENNETT, D.C. (1984) Elbow Room: Varieties of Free Will Worth Wanting. MIT Press.

DRETSKE, F. (1981) *Knowledge and the Flow of Information.* MIT Press.

DRETSKE, F. (1988) *Explaining Behavior. Reasons in a World of Causes.* MIT Press.

DRETSKE, F. (1995-a) *Naturalizing the mind.* MIT Press.

DRETSKE, F. (1995-b) Does meaning matter? in C. Macdonald and G. Macdonald (1995)

DRETSKE, F. (2006) Representation, teleosemantics and the problem of self-knowledge. in G. MACDONALD & D. PAPINEAU (2006-b)

EGAN, F. (1991) Must psychology be individualistic? *Philosophical Review* 100: 179-203.

EVANS, G. (1982) *The Varieties of Reference.* Oxford University Press.

FARKAS, K. (2003) What is externalism? *Philosophical Studies* 112/3: 187-208.

FARKAS, K. (2008) *The Subject's Point of View.* Oxford University Press.

FIELD, H. (1978) Mental representation. *Erkenntnis* 13: 9-61.

FODOR, J.A. (1975) *The Language of Thought.* MIT Press.

FODOR, J.A. (1987) *Psychosemantics.* MIT Press.

FODOR, J.A. (1991) A modal argument for narrow content. *Journal of Philosophy* 88: 5-26. Reprinted in Macdonald and Macdonald (1995)

FODOR, J.A. (1994) *The Elm and the Expert: Mentalese and its Semantics.* MIT Press.

GERTLER, B. (2007) Content externalism and the epistemic conception of the self. *Philosophical Issues* 17: 37-56.

GRUSH, R. (2001) The semantic challenge to cognitive neuroscience. in P. Machamer, P. McLaughlin & R. Grush (eds.) *Theory and Method in the Neurosciences.* University of Pittsburgh Press.

HADDOCK, A. & MACPHERSON, F. (eds.) (2008) *Disjunctivism: Perception, Action, Knowledge.* Oxford University Press.

HEIL, J. (2004) Natural intentionality. In R. Schantz (2004)

HEIL, J. & MELE, A. (eds.) (1993) *Mental Causation*. Oxford: Clarendon Press.

HORGAN, T.E. (2001) Causal compatibilism and the exclusion problem. *Theoria* 16: (40) 95:116.

HORGAN, T.E. & TIENSON, J. (2002) The phenomenology of intentionality and the intentionality of phenomenology. in D. Chalmers (ed.) *Philosophy of Mind.* Oxford University Press.

HUMPHREYS, P. (1999) Causation. in H.W. Newton-Smith (ed.) *A Companion to Philosophy of Science.* Cambridge, Mass.: Blackwell.

JACKSON, F. (1998) Reference and description revisited. *Philosophical Perspectives* 12, 201–218.

JACKSON, F. & PETTIT, P. (1988) Functionalism and broad content. *Mind* 97: 381-400.

JACKSON, F. & PETTIT, P. (1990) Program explanation: A general perspective. *Analysis*, vol.50, no.2: 107-117.

JACOB, P. (1997) *What Minds Can Do: Intentionality in a Non-Intentional World.* Cambridge University Press.

KIM, J. (1989) Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives* 3: 77-108. Reprinted in Kim 1993, pp. 237-64.

KIM, J. (1993) *Supervenience and Mind.: Selected Philosophical Essays.* Cambridge University Press.

KIM, J. (1998) *Mind in a Physical World*. MIT Press.

KIM, J. (2005) *Physicalism, or Something Near Enough.* Princeton University Press.

KRIEGEL, U. & HORGAN, T. (forthcoming) The phenomenal intentionality research program. In U. Kriegel & T. Horgan (eds.) *Phenomenal Intentionality.* Oxford University Press.

KRIPKE, S. (1972) *Naming and Necessity***.** Oxford: Blackwell.

KROON, F. (1987) Causal descriptivism. *Australasian Journal of Philosophy* 65: 1–17.

LEWIS, D. (1984) Putnam's paradox. *Australasian Journal of Philosophy* 62: 221–36.

LOAR, B. (1988) Social content and psychological content. In R. Grimm and D. Merrill (eds.) *Contents of Thought.* University of Arizona Press.

LOEWER, B. (1987) From Information to intentionality. *Synthese* 70: 287-317.

MACDONALD, C. & MACDONALD, G. (eds.) (1995) *Philosophy of Psychology: Debates on Psychological Explanation.* Oxford: Blackwell.

MACDONALD, G. & PAPINEAU, D. (2006-a) Introduction: Prospects and problems for teleosemantics. in G. Macdonald & D. Papineau (2006-b)

MACDONALD, G. & PAPINEAU, D. (2006-b) *Teleosemantics.* Oxford University Press.

MALCOLM, N. (1968) The conceivability of mechanism. *Philosophical Review* 77: 45-72.

MCDOWELL, J. (1977) The sense and reference of a proper name. *Mind* 86: 159-185.

McDowell, J. (1994) *Mind and World.* Cambridge, Mass.: Harvard University Press.

MENZIES, P. & PRICE, H. (1993) Causation as a secondary quality. *British Journal for the Philosophy of Science* 44: 187–203.

MILLIKAN, R. (1984) *Language, Thought and Other Biological Categories.* MIT Press.

MILLIKAN, R. (1993) *White Queen Psychology and Other Essays for Alice.* MIT Press.

MILLIKAN, R. (2004-a) Existence proof for a viable externalism. in R. Schantz (2004)

MILLIKAN, R. (2004-b) Comments on "Millikan's (Un?)Compromised Externalism". In R. Schantz (2004)

NEWMAN, A.E. (2003) *Causal Efficacy and Externalist Mental Content.* Unpublished doctoral thesis. http://hdl.handle.net/1721.1/8147

O'BRIEN, G. & OPIE, J. (2004) Notes towards a structuralist theory of mental representation. In H. Claplin, P. Staines & P. Slezak (eds.) *Representation in Mind: New Approaches to Mental Representation.* Amsterdam: Elsevier.

PAPINEAU, D. (1993) *Philosophical Naturalism.* Oxford: Blackwell.

PESSIN, A. & GOLDBERG, S. (eds.) (1996) *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'".* New York: M.E. Sharp.

PLANTINGA, A. (1993) *Warrant and Proper Function.* Oxford University Press.

PUTNAM, H. (1975) The meaning of 'meaning'. in H. Putnam, *Mind, Language and Reality: Philosophical Essays. Philosophical Papers Vol.2.* Cambridge University Press.

PUTNAM, H. (1997) Functionalism: Cognitive science or science fiction. in D.M. Johnson and C.E. Erneling (eds.) *The Future of the Cognitive Revolution.* Oxford University Press.

RAATIKAINEN, P. (2010) Causation, exclusion and the special sciences. *Erkenntnis* 73 (3): pp. 349-363.

REY, G. (1997) *Contemporary Philosophy of Mind: A Contentiously Classical Approach.* Cambridge, Mass.: Wiley-Blackwell.

RYLE, G. (1949/1984) *The Concept of Mind.* The University of Chicago Press.

SCHANTZ, R. (ed) (2004) *The Externalist Challenge.* New York: Walter de Gruyter.

SEARLE, J.R. (1980) Minds, brains and programs. *Behavioral and Brain Sciences* 3: 417-424.

SEARLE, J.R. (1983) *Intentionality.* Cambridge University Press.

SEGAL, G. (1989) Seeing what is not there. *Philosophical Review* 98: 189-214.

SEGAL, G. (2000) *A Slim Book about Narrow Content.* MIT Press.

SHOEMAKER, S. (2007) *Physical Realization.* Oxford University Press.

STALNAKER, R. (1989) On what's in the head. *Philosophical Perspectives* 3: 287-319. Reprinted in Pessin and Goldberg (1996)

STICH, S. (1983) *From Folk Psychology to Cognitive Science: The Case Against Belief.* MIT Press.

TURING, A. M. (1950) Computing machinery and intelligence. *Mind* 50: 433-460.

TYE, M. (1995) *Ten Problems of Consciousness.* MIT Press.

WALTER, S. (2009) Taking realization seriously: No cure for epiphobia. *Philosophical Studies.* Vol. 151/2: 207-226

WOODWARD, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

YABLO, S. (1992) Mental causation. *Philosophical Review* 101: 245-80.