University of Tartu

Institute of Philosophy and Semiotics

ON THE IMPLAUSIBILITY OF SLOW-SWITCHING ARGUMENTS IN ESTABLISHING INCOMPATIBILITY THESIS

Master's Thesis in Philosophy

Arunkumar Rajavel

Supervisor(s):

Prof Bryan Frances

Prof Juhani Yli-Vakkuri

Tartu

2019

TABLE OF CONTENTS:

1. INTRODUCTION

Philosophical enquiry can often surprise us by arriving at conclusions (or dilemmas) that run counter to our intuitions. The relationship between content externalism and authoritative self-knowledge is one of such areas. While there are strong reasons to think that content externalism is true, an even stronger intuition is at play that motivates us to hold on to the notion of us having privileged access to our own thought contents. After all, the very suggestion that we may not have an authoritative self-knowledge seems so alien to us. Any philosopher who wishes to subscribe to both the stances will find themselves in an impasse for reasons that incompatibilists point out. This thesis is an attempt to offer solutions to such philosophers.

The structure of this thesis is as follows: with this chapter providing some introduction to the thesis, laying out what this thesis is, and is not, about, in chapter 2 and its sections, I provide an overview of the twin-earth thought experiments that birthed this whole debate of the incompatibility thesis. In chapter 3, I introduce variants of slow-switching arguments and argue against both the variants, namely the content switch and the memory argument. In section 3.1, I argue against—what I call—the content switch argument and, in section 3.2, against the memory argument. I divide section 3.2 into two subsections each of which advances a separate strategy to attack premise 3 of the memory argument. Having argued against both the variants of slow-switching cases, I proceed to conclude that slow-switching cases do not succeed in establishing the incompatibility thesis in chapter 5. But before that, in chapter 4, I take on four potential objections to my arguments.

A few preliminaries before moving on:

- I shall not be concerned with establishing the truth of either content externalism or of authoritative self-knowledge. In the context of this debate it is standardly assumed that externalism is true and that, if it is true, then it is not compatible with us having authoritative self-knowledge.
- In line with much of literature, this essay is concerned only with content externalism and not other kinds of externalism. Sometimes I use "externalism" as a shorthand for "content externalism." For an overview of different types of externalism, please see (Carter et al., 2014). Also, as I'm only concerned with

authoritative self-knowledge, the terms "thought content(s)" or "mental content(s)" refers only to "introspective thought content(s)," wherever applicable.

- Recent empirical advances suggest that the commonly accepted intuitions about twin earth scenarios may, after all, not be as common as it was purported; the relevant intuitions vary with respect to culture and the language used. For a discussions on this, see (Tobia et al., 2017). This might be the case but I do not concern myself with this worry and I situate my work wholly within the traditional context in which this debate operates—that is, the English-speaking world.
- I do not take it upon myself to define what I mean by "external environment" (or, for that matter, "internal environment"). Considerable disagreements prevail on what an external environment is but, as it does not have any effect on the arguments of my thesis, I do not define it. The reader may understand it in any way that they deem fit.
- As will become apparent below, mental concepts play a key role in my thesis and thus cases concerning empty concepts (such as that of Boghossian's Dry Earth cases) are beyond the scope of this thesis[1]. Although Boghossian himself argues that such cases do not threaten the status quo of the incompatibility debate (whichever side you subscribe to), in this thesis I do not consider such cases.
- In line with the standard notation employed in cognitive sciences literature, I use capitalised words for concepts. For example, 'WATER' stands for the concept of water, and 'water' stands for the extension of the concerned object, i.e. water itself.
- Finally, I'm not concerned with self-knowledge simpliciter, but with authoritative self-knowledge (also called "privileged access" or "first-person authority"). In this thesis, I occasionally use the term "self-knowledge" alone, but throughout the text, wherever applicable, it must be understood as though I'm talking about authoritative self-knowledge.

When I say "concepts," I mean the things that one's thought expresses. It is akin to the meaning or the content expressed by a linguistic expression. When I say "reference," I

[1] For a treatment of related issues, see (Boghossian, 1997).

mean the object out there that the thought (uniquely) picks out[2]. It is akin to the unique object that a linguistic expression picks out. Concepts and references are to mental contents as connotations and denotations (in Millian terminology) are to linguistic expressions; or, to use Carnap's terminology, intension and extension respectively.

With recent experimental evidences from psychological research questioning the very idea of our having privileged access to our thoughts, I do not endorse any philosophical stance towards the possibility of having, or not having, authoritative self-knowledge in this thesis. I merely argue against the plausibility of slow-switching arguments of the incompatibility thesis which holds that if content externalism is true, then it is incompatible with authoritative self-knowledge. I conclude that slow-switching arguments do not satisfactorily establish the incompatibility thesis.

[2] Boghossian (1997) distinguishes a referent from its extension and this distinction would take us to the highly muddled discussions on natural kinds being rigid designators and, into essentialism. Since this distinction, or lack thereof, does not upset my arguments, I take it to the case that a term's reference and its extension are the same. For a discussion of related issues, please see (Ben-Yami, 2001).

2. TWIN EARTH THOUGHT EXPERIMENTS—A SUMMARY

In this chapter, I furnish the required background for the incompatibility thesis. Below, in section 2.1, I provide a brief introduction to Putnam's twin earth thought experiment which concerns the role of external physical environment in determining the meaning of linguistic expressions; and Burge's extension of this thesis to incorporate, in addition to external physical environment, the influence of socio-linguistic factors. I set aside section 2.2 to elucidate the incompatibility thesis, in which I pay specific attention to slow-switching thought experiment, the subject of my thesis. The reader who is well-versed in these topics can skip the entirety of this chapter, and jump to chapter 3.

2.1 PUTNAM'S AND BURGE'S THOUGHT EXPERIMENTS

Putnam was one of the myriad philosophers who attempted to answer the question, "what gives words their meanings?" The Fregean notion, which still remains as one of the central theses in this research area, was that meaning corresponds to the mental states of the agents. That is, if two words have same meaning then they must have the same reference. Or, in other words, two words with different references cannot possibly have the same meaning. Putnam found this contentious and argued that meaning is imparted, at least, partially, by the external environment, challenging the Fregean notion. In effect, he put forth the thesis of semantic externalism, the idea that meanings of words (or any linguistic expression) are individuated, at least, partially, by factors external to the agent. His argument involves a thought experiment which later came to be dubbed the twin earth thought experiment (Putnam, 1974).

Imagine a distant planet in another galaxy or another universe that is exactly the same as that of earth in all respects—where there is a molecule-to-molecule physical duplicate for everything that is on earth. The only difference between earth and this "twin earth" (henceforth shall be called twin earth) is that while water is composed of molecules of H_2O, the twin water ("twater" for brevity) on twin earth is made of up of some unknown compound, say, XYZ. However, the thought experiment is set in a time when we haven't discovered yet that water is made up of H_2O and thus, propositions like "water is H_2O" are not yet conceivable by the agents. Now consider an individual called Oscar and his twin,

who is a molecule-to-molecule duplicate of Oscar, called Toscar (he is also called "Oscar" but we use "Toscar" for the purposes of differentiation). When Oscar says "water is wet," he will be referring to the water on earth, i.e. samples of H_2O molecules. Whereas if Toscar says, "twater is wet," he will be referring to twater on twin earth, i.e. samples of XYZ molecules. Having never had any contact with the other alternative, both Oscar's and Toscar's usage can only be directed at their immediate surroundings. If meanings are a reflection of internal mental states of the agents, then it is not possible for two molecule-to-molecule duplicates like Oscar and Toscar to say the same word and mean different things. Thus, Putnam argues, meanings are not in the head. A word of caution is in need. Putnam's usage of a natural kind term like water has many unfortunate consequence such as Oscar and Toscar not being able to an exact molecule-to-molecule duplicate as Oscar's body will be made up of H_2O while Toscar's that of XYZ; or, that the unknown effect that twater may have on Oscar's body which was used to water; and so on. But for the purposes of argumentation let us set aside these technical deficiencies. These do not actually undermine the twin earth thought experience, for Putnam's other examples such as those of elms and beeches do not face this problem. Withal, in line with majority of the literature, I take up only water/twater cases in this thesis.

Although Putnam propounded externalism for language, the basic notion of externalism is in use in a wide range of philosophical domains such as epistemology, historiography of science, and moral philosophy. In the philosophy of mind, it is a thesis that holds that contents of our thoughts are individuated by the relational properties between the mind and the world. Tyler Burge famously expanded Putnam's thought experiment (which concerns only the physical environment) to demonstrate that sociolinguistic factors individuate our mental contents as well (Burge, 1979). For instance, factors such as language, social conventions, rituals, etc., also determine one's thought contents.

Suppose that Oscar is a normal adult living in an English-speaking country on earth where the term "arthritis" is used to talk about the disease that affects the joints. Imagine Oscar's molecule-to-molecule duplicate, Toscar, living in a twin community (on twin earth) where the term "tharthritis" (twin arthritis, pronounced the same like arthritis, but is written as "tharthritis" for purposes of differentiation) is used to talk about the disease that

affects both the joints and the muscles. Neither Oscar nor Toscar are medical experts and do not know the exact applicability conditions of the term. Now suppose that upon observing an inflammation on his thighs, Oscar thinks to himself: "I have arthritis on my thigh." If he goes to his doctor and tells the doctor that he has arthritis on his thigh, the doctor will correct Oscar that arthritis is a disease that can only affect the joints and not the thighs. Thus, Oscar realises that he has a false belief. Now suppose that under similar circumstances on twin earth, Toscar exclaims to his (twin) doctor that he has tharthritis on his thighs. As in this community, "tharthritis" is a term that is applied to both the muscles and the joints, the twin doctor proceeds to treat Toscar. In this case Toscar has a true belief. Through this example, Burge argues that if contents of our minds are dependent only on the inner states of the agent, it is not possible for two molecule-to-molecule duplicates like Oscar and Toscar to utter the same sentence (or hold the same belief) and yet differ in truth values—as is the case here where one holds a true belief and the other false belief. In this manner, Putnam and Burge forward the externalist theses.

Thus, externalism (content externalism) in the philosophy of mind is the thesis that the content or meaning of a thought is dependent, at least, partly on the external environment—be it physical or social (linguistic community, for example). That is, contents of one's thoughts are, at least, partly individuated by factors that are external to the agent. Internalism, on the other hand, denies this position and maintains that contents or meanings of thoughts supervene only on the inner states, or, in other words, on the intrinsic properties, of the agent. "Intrinsic properties" need not necessarily only concern standalone properties that are intrinsic to the agent; it may also concern relational aspects that are not outside the agent's body. For example, two-place relations among neurons can be part of the base that determines mental contents. These are relations, instead of properties, but still count as "intrinsic to the agent."

Thus, in this section I outlined Putnam's and Burge's thought experiments which provides the broader context on what externalism is. In the next section, I sketch a short introduction to incompatibility thesis.

2.2 INCOMPATIBILITY THESIS

In the preceding section, we saw the "bigger picture" background of this whole internalism/externalism debate. In this section, I delineate how the above two thought experiments gave rise to the incompatibility thesis—the notion that content externalism is incompatible with authoritative self-knowledge.

As we saw, content externalism in the philosophy of mind is the thesis that the content or meaning of a thought is dependent, at least, partly on the external environment—be it physical or socio-linguistic. An unfortunate consequence of externalism is that, if externalism is true,—that is, if our thoughts are individuated by relational aspects between the environment external and the agent—then that entails that we cannot know our own thoughts without launching an empirical investigation of our environment. However, we have this traditional (at least, since Descartes), and often intuitive, notion that we have a certain privileged access to our own thoughts; that we can know our own thoughts without the need to investigate our environment; that we can know our own minds in more authoritative, and infallible (this is controversial), a way than we can know others' minds. This notion is called authoritative self-knowledge. So one must either renounce their subscription to externalism or to us having authoritative self-knowledge. Or, in other words, these two are incompatible with each other. The idea that externalism is incompatible with our privileged access to our own thought contents is first put forth by Burge invoking, and expanding, Putnam's twin earth thought experiment (Putnam, 1974; Burge, 1979; 1988).

This observation inspired a flurry of activity in the philosophical circles. Paul Boghossian assesses Burgean worries and concludes that externalism is not compatible with self-knowledge (Boghossian, 1989). This paper, indeed, argues that given externalism, none of the available theoretical options to explain us having self-knowledge—namely our having privileged access to our minds through inference, or on the basis of introspection, or on the basis of nothing—can satisfactorily provide an account of authoritative self-knowledge and concludes that given externalism, we cannot have privileged access to our mental contents. Ever since the publication of Burge's and Boghossian's influential papers, the philosophical community has been actively debating

these issues. There are, in effect, two positions on the relationship between content externalism and authoritative self-knowledge: the thesis that they are compatible is called compatibilism, and the thesis that denies it is called incompatibilism. Most externalists belong to the former camp while most internalists to the latter (but not necessarily so).

In this short section, I provided a short overview on the incompatibility thesis. The incompatibility thesis is advanced through two strategies namely the reductio ad absurdum and slow switching arguments of which I'm arguing only against the plausibility of, as the title suggests, the latter. In the next chapter, I provide my arguments against both the variants of the slow-switching argument.

3. AGAINST THE SLOW-SWITCHING ARGUMENTS

In section 2.2, I provided an overall background to the incompatibility thesis. The incompatibility thesis is forwarded by two class of arguments that follow different strategies. They are as follows:

A. **Reductio-type argument,** due to Michael McKinsey, reasons that if externalism is true, it leads to the implausible conclusion that we can know *a priori* certain things about our environment that can only be known *a posteriori* (McKinsey, 1991; Warfield, 1992). It can be rendered as follows:

   1. If Oscar thinks that water is wet, then water exists.
   2. Oscar thinks that water is wet.
   3. Therefore, water exists.

   The idea is that if content externalism is true, (1) is true. (3) is an empirical proposition that cannot be known *a priori* but only upon empirical investigation. But our accepting (1) and (2) logically entails (3); that is, it indicates that (3) can be known *a priori*, which is not true. Thus, McKinsey argues, accepting content-externalism leads to counterintuitive conclusions and is, therefore, incompatible with authoritative self-knowledge. As this line of argumentation is beyond the scope of this essay, I do not concern myself with this. For criticisms/commentary on this, see (Boghossian, 1997; Brueckner, 1992; Pryor, 2007) to name a few.

B. **Slow-switching argument**, the second strategy, judging from the vast amount of literature it inspired, could be considered as the most influential argument for the thesis of incompatibilism. Suppose that Oscar from Putnam's original twin earth thought experiment is transported to twin earth without his knowledge in some way: say, he was secretly transported to twin earth while he was asleep and he wakes up ignorant of the switch. He is let to live there on twin earth for a very long time. Since the state of affairs on twin earth is an exact replica of everything on earth, save for the molecular composition of twater, he does not notice any difference and goes on about his every day business as usual. This is termed as

> ‘slow-switching.’ The incompatibilists reason that, in such a slow-switching scenario, the contents of Oscar’s thoughts switch gradually from being water thoughts to twater thoughts unbeknownst to him and at a certain point will become fully twater thoughts. Suppose we tell Oscar about the switch later (without specifying *when* the switch occurred) and ask him to distinguish which of his thoughts were water thoughts and which were twater thoughts, he would be unable to differentiate. As Burge puts it: “...the person would have different thoughts under the switches, but the person would not be able to compare the situations and note when and where the differences occurred.” (Burge, 1988, pp.653).

Incompatibilitists advance this argument to argue that in a slow-switched scenario, since Oscar can neither **differentiate the switch** in his mental contents before he was made cognisant of the switch, nor spot **when** his thought contents switched after he was being told of the switch, he does not know his own thoughts—that is, he lacks authoritative/privileged access to his own mental contents. Therefore, they conclude, if content externalism is true, it is incompatible with the doctrine of authoritative self-knowledge. A number of philosophers have attempted in a variety of ways to argue against this thesis (Falvey & Owens, 1994; Warfield, 1992; 1997; Gibbons, 1996; Morvarid, 2015; Parent, 2015).

The slow-switching argument, in itself, has two variants. I dub the aforementioned argument the “content-switch argument,” and a variant of this parent argument is called the “memory argument” which first appeared in (Boghossian, 1989). In the following sections, I argue against both the variants of the slow-switching arguments with special emphasis on the latter. I provide one argument against the content switch argument and two arguments against the memory argument. The first of the two arguments I advance against the memory argument is an extension of my reasons to reject the content switch argument—thus, it is, in effect, one argument covering both the variants—and the second one is an independent strategy. Through my argumentation in the sections of this chapter, I aim to demonstrate that neither of the variants of slow-switching arguments succeed in establishing the incompatibility thesis.

3.1 CONTENT SWITCH ARGUMENT

Much of the literature in this debate holds that thought contents of a slow-switched agent (Oscar) change unbeknownst to the agent provided the agent stays there long enough. Although no exact figure had been arrived at, it is generally held that it will take years together for this switch to occur (Brueckner, 1997). If Oscar was transported to twin earth and he spends a long time on twin earth while causally interacting with twin-earthian objects, his thoughts and linguistic expressions would gradually come to express twin-earthian concepts and refer to twin-earthian objects. This much is in consensus among both the internalists and externalists (Burge, 1989; Ludlow, 1995a, 1995b; Smith, 2003; Vahid, 2003). I deny that conceptual switches take place.

Despite the fact that it was not explicitly acknowledged in the literature, I interpret this notion of automatic content switch to implicitly presuppose that concept nativism is true. If concept nativism is true, then it is possible for us to acquire concepts automatically without us knowing. Before I address this, let us take a quick look at what concept nativism and, its opposing position, concept empiricism are.

There are two theses concerning the innateness of concepts, namely concept nativism and concept empiricism, and depending upon which conceptual structure one subscribes to (such as the classical theory of concepts, prototype theory, theory theory and conceptual atomism), one has to commit to varying levels of innateness of concepts (for an overview of the different conceptual structures, please refer (Laurence and Margolis, 1999)). Nativistic accounts of mental concepts posit that most of our mental concepts are innate and that they are lying about in wait for appropriate stimuli to emerge and "activate" them for the agent to grasp[3]. While this may appear counterintuitive, there are strong reasons that motivates subscription to such a view. As Chomsky argues, the only explanation that can account for the puzzle of our language acquisition would be that our minds must have certain innate language learning capabilities. As it was dubbed, there

[3] We are not concerned with the extreme position endorsed by Jerry Fodor called "radical concept nativism" which holds that *all* of our concepts are innate and that, no concept can ever be learned. Although Fodor had had independent reasons to arrive at this unintuitive and disheartening conclusion, this extreme view has been severely discredited (Margolis, 1998). A form of concept nativism, which holds that there are many innate concepts that cannot be learned, survives from this proposal and is prominent in mainstream contemporary debate on this topic.

seems to be a "poverty of stimulus,"—an impoverished stimuli,—that still does not deter a child from acquiring language (Laurence and Margolis, 2001). Concept nativism in the philosophy of mind is simply Chomsky's supposition that there must be innate language learning capacities in children extended to mental concepts. All of the aforementioned conceptual structures commit to the existence of innate concepts with the commitment of concept atomism being the most; one can think of atomism as a form of radical concept nativism.

Concept empiricism, on the other hand, holds that there are very few, if any, concepts that are innate (which are dubbed "primitive concepts") and that most other complex concepts must be acquired through empirical means through our interactions with the world. Concept empiricism does solve many of the issues that concept nativism faces but also faces certain shortcomings. Again, although the debate whether our conceptual system is innate or empirical is unsettled, learning plays a key, if not the only, role in the acquisition of new concepts. Both the nativists and empiricists agree that learning a new concept involves learning and only differ in which concepts are innate and the way (mechanism) they can be learned.

The standard reading of concept atomism renders it on par with extreme concept nativism. That is, it was usually held that learning has no role in conceptual atomism but (Margolis, 1998) argues that there is a space for learning in this framework, too. For the purposes of my thesis, it wouldn't matter which conceptual structures I favour.

Now to return. To say that thought contents change through continued, causal interaction with twater is akin to saying when an appropriate stimuli such as twater is pressed forth, the concept TWATER emerges for Oscar to grasp. Or, in other words, there innately is, in Oscar's conceptual system, a sustaining mechanism that will get triggered with the presence of twater thereby letting Oscar grasp the concept TWATER without him having to learn anything new. In concept nativism, the following holds:

> "…a sustaining mechanism [...] is all wired up in advance and simply waiting for an innately specified triggering condition to cause it to become activated. Far from it. What is innate, according to the model, is a general cognitive organization for creating a range of syndrome-based sustaining mechanisms in response to new natural kinds." (Laurence and Margolis, 2011, pp.524).

But it is also the case that all the models of conceptual structure, including the classical theory, hold that there are certain innate concepts (primitive) that cannot be learned. If natural kinds turn out to be innate in one's mind, then that would provide the incompatibilist an edge, for if that were the case, Oscar need not put in conscious efforts to learn of twater. Margolis and Laurence, themselves nativists, provide a list of concepts that are taken to be innate according to contemporary concept nativists. They say:

> "Likely candidates for innate concepts include concepts associated with objects, causality, space, time, and number, concepts associated with goals, functions, agency, and meta-cognitive thinking, basic logical concepts, concepts associated with movement, direction, events, and manner of change, and concepts associated with predators, prey, food, danger, sex, kinship, status, dominance, norms, and morality." (Margolis and Laurence, 2011, footnote: 21).

We can see that this list does not include natural kinds. Notwithstanding that, let us grant that natural kinds are unlearnable primitive concepts which are innate in our mind's conceptual system. (Margolis and Laurence, 2011), as a response to Fodor's radical concept nativism (see footnote no.2), surveys various methods of concept acquisition and argue that all concepts, including primitive concepts, can be learned.

> "...Though the distinctive character of nativist and empiricist accounts of concept learning differ, for nativists, just as for empiricists, learning is absolutely central to the explanation of concept acquisition. The burden of this paper has been to show that the commitment to learning that both sides share is perfectly cogent." (Margolis and Laurence, 2011, pp. 538).

Thus, so far I argued for the following:

i) The classical theory of concepts hold that certain primitive concepts are unlearnable and are, thus, innate. If water/twater concepts are one such, then Oscar need not learn the twin concepts and that he will automatically acquire them.

ii) Then we saw that the list of concepts considered primitive do not include natural kinds and thus natural kind concepts must be learned.

iii) Finally I reasoned that even if we are wrong about this, primitive concepts still should be learned somehow.

So I take it to be settled that learning a concept is inevitable for our acquiring new concepts. But through this, I do not endorse that acquiring a new concept have always to be active; that passive, or semi-/unconscious learning, is not possible. Far from that. If that were true then that goes against the very idea of content externalism. What I claim is that to be granted for our thought contents to change, some form of new concept must be learned, where the learning could be semi-/unconscious. In other cases (such as when we first encounter a new concept) it could well be semi-/unconscious but the only way in cases such as water/twater, arthritis/tharthritis that we can learn the new twin concept is through learning that one crucial difference that separates the two. There is no other way how Oscar would come to know of twater[4]. Similarly, for Twin Oscar (Toscar), his default concept would be that of twater and his slow switching to earth would require that he learns how to distinguish water and twater.

To take another example: consider how certain words, through passage of time, change their meaning. An agent who lives through this change can be thought of being slow switched. The critic might say that the agent's hought contents would have, along with his deference to community for correct usage, slowly changed along with the changes in the meaning of the word. For this purpose, we can assume that the agent was in comatose for, say, 100 years and wakes up without any idea that the meaning of the word they once knew has completely changed (assume that he gets to live a very long life). When this agent wakes up and goes about interacting with others, using terms such as "gay," (which only meant happiness before he went to coma and now almost always means homosexuality), "salad," (whose meaning has changed to accommodate inclusion of meat parts, too, in the mix) and so on. Unless this agent runs into severe repercussion which results in him learning of the changes and thereby updating/changing his concepts GAY and SALAD, he will continue to use them in only the way that *he* intended. If the community were to be so understanding that they don't bother to correct him or educate him and communicate accordingly with him, there is no way his thought contents would change. Thus, in this example unless the agent were to make a conscious effort to distinguish between the new twin concepts, say, $SALAD_1$ and $SALAD_2$, the agent's

[4] One can, of course, ask: what about possessing concepts through deference, then? I address this objection later in this thesis. Please refer section 4.1.

thought contents will not change. A similar example is provided in section 4.1 for ARTHRITIS and THARTHRITIS.

Thus, no matter what the conceptual scheme of our mind is, in cases like the slow-switching scenario, it is absolutely crucial for Oscar to learn the concept TWATER in order to acquire it. In the event that Oscar does not make active efforts to learn the concept TWATER, he shall never come to acquire it, and thus, in his interactions with the twin-earthians, his usage of "water" will express in his mind the concept WATER, whereas in the twin-earthians' usage of the same word will express the concept TWATER[5]. In the event that Oscar engages in active cognitive participation to learn the new concept TWATER,—say by learning the one unique feature that distinguishes water and twater: the molecular structure—he would come to know the change in his mental concepts. Therefore it cannot be claimed that in either of the cases—before and after learning the new concept TWATER—he does not know his own thought contents.

Now if all of this fails to convince a critic, consider the following case: it was said that Oscar has no clue about molecular structures and so on. Let us suppose that Bob is a chemist who has been slow-switched. If asked to explicate what he means by water, he would provide a list of chemical properties of water (all of which are shared by twater) and will also say that it is made up of molecules of H_2O. To this, a twin-earthian chemist would tell Bob that twater is made up of XYZ. Now I see no reason how and why Bob who is a chemist, whose concept of water (H_2O molecules) would automatically switch from being WATER to TWATER without him learning the one crucial difference that separates the two. Simply put, I just maintain that, not only for the slow-switched chemist Bob, but also for the chemically-ignorant Oscar, the only way their concepts change—in cases similar to those discussed in slow switching scenarios—is for them to make active efforts to learn the difference between the two concepts.

One seemingly severe objection that could be raised can be phrased as follows: by granting that an agent's awareness is needed for thought contents to change, don't you downplay that an agent can come to possess certain concepts through semi-/unconscious means, through deference, and so on? Does this not go against the very thesis of content

[5] This supposition may give rise to an objection that it signifies that essentially a miscommunication is taking place. I discuss this objection in section 4.3.

externalism? This is a valid criticism but I maintain that this does not necessarily upset externalism. My claim is that only in cases that slow-switching arguments deal with the content switch does not happen—that is, two mutually indistinguishable environments which have only one key difference that is unknown to the agent. This does not, in any way, downplay the influence of the environment on individuating Oscar's thought contents. The new environment does exert and influence the occurrent thought contents (memorial contents are a different story and I deal with it in the following section) of the slow-switched agent. The agent has his thought contents informed by the new environment and in the case of the agent not knowing the crucial differences, as is the case now, the agent would only be influenced by factors from the new environment that are also shared by the old environment; and thus 'picks up' the same factors that end up determining the agent's mental contents. In the absence of having learnt the new concept, the agent, having 'picked up' the same old factors from the present environment, will only be able to think of the old concept. This becomes even clearer when we consider the case of slow-switched Dry Oscar (Doscar). Please refer section 4.1 for further details.

Now, to summarise: assuming the truth of externalism, my arguments against content switches can be presented as follows:

1. If content externalism is incompatible with authoritative self-knowledge, then in a slow switched situation, Oscar's occurrent thought contents can, owing to the new environment's influence on his mental contents, change unknown to him.
2. Our best theories of concept acquisition mandate that all concepts, no matter how primitive, are to be learned, and that there are no concepts that change/gets added to our conceptual repertoire without us acquiring it through learning.
   2.1. Learning can be semi-/unconscious.
3. Possession of any concept, thus, requires some form of learning.
4. We can grant that Oscar possesses both WATER and TWATER concepts only if he can distinguish between the two, in the event that all the n-1 properties are shared between water and twater.
5. The only way Oscar can distinguish between WATER and TWATER is by learning the corresponding differences in their molecular composition.
6. Before being told of the switch, Oscar does not learn the molecular compositions and thus, cannot distinguish between WATER and TWATER concepts.

6.1. If he was told of the switch, Oscar will learn the differences in molecular composition and can, thus, distinguish between WATER and TWATER.

6.2. In this case, he will be able to know that his thought contents have changed (thus, his thought contents change *not unknown* to him).

7. By 5 and 6, we cannot grant that Oscar possesses both WATER and TWATER concepts; and so is left with his old WATER concept alone, using which he keeps miscommunicating with the twin earthians.
8. By 7, we can ascertain that Oscar's occurrent thought contents do not change unknown to him in both the cases where he was, and he was not, told of the switch.
9. Thus, by 8 and 1 (modus tollens), it is not the case that content externalism is incompatible with authoritative self-knowledge.

Since as I argued above, the thought contents of a slow-switched agent do not change unknown to the agent, slow-switching argument does not establish incompatibility thesis.

3.2 MEMORY ARGUMENT

The memory argument is a variant of the slow-switching argument. Commenting on the Burgean notion of occurrent thoughts having a self-verifying status, Boghossian argues that content externalism, if applied to contents of one's memories, would be incompatible with self-knowledge (Burge, 1988; Boghossian, 1989). The conclusion of Boghossian's argument in this paper can be also thought of as the sceptic's argument to self-knowledge. The argument, as rendered by Peter Ludlow, is as follows (Ludlow, 1995b):

1) If S forgets nothing, then what S knows at t1, S knows at t2.
2) S forgot nothing.
3) S does not know that *p* at t2.
4) therefore, S did not know that *p* at t1.

Naturally, the memory argument has seen a variety of responses. For example, Sven Bernecker writes that this formulation is built upon the assumption that presentist externalism is true, but that when it comes to memorial contents, only pastist externalism applies; and that with this, memory argument does not succeed (Bernecker, 2004).

Anthony Brueckner, on the other hand, rejects premise 2 by arguing that not forgetting something does not imply having an intact knowledge of the past; that one can lose knowledge in other ways besides forgetting, for example, in the event of a new defeating condition (Brueckner, 1997). Curiously, others like Sanford Goldberg even suggest that the memory argument can be forwarded without involving the faculty of memory at all and that the same conclusion can be reached by having the agent repeat a verbalised thought all the time (Goldberg, 1997). For a brief survey of various attacks on the above formulation, see (Ludlow, 1999). Also see (Falvey, 2003).

In addition to these aforementioned challenges, I take upon myself to argue against the memory argument through two strategies, where both deny premise 3 on different grounds. The first one of them is regarding the acquisition of concepts, similar to the one I gave in 4.1, applied to the memory argument. The second strategy is based on how memory actually works and that, all things kept equal, premise 3 is false. Let us see the first of my argument below.

3.2.1 DENIAL OF CONTENT SWITCH FOR MEMORY

In this subsection, I deny premise 3 directly by appealing to my argument in section 3.1 that concept—and, thus, consequently, the reference—switch is not automatic. The memory argument, too, just like the content switch argument, implicitly assumes that Oscar's thought contents have changed from being water thoughts to twater thoughts without him knowing it, and when Oscar remembers what he thought of at t1 (some particular water thought), he actually 'remembers' twater thoughts at t2. In Ludlow's formulation above in 3.2, *p* is a reflexive proposition of the form, "S thinks that *q*," in which *q* is a first-order proposition. In our case, let *p* be "S thinks that water is wet," in which *q* stands for the first-order proposition, "water is wet." And, on twin earth, correspondingly, let *p'* stand for "S thinks that twater is wet," and *q'*, respectively, for "twater is wet." Now, according to the memory argument, Oscar thought that *p* at t1 and when he remembers this thought at t2, he remembers *p'* as his thought contents have changed. If we grant that Oscar knew that *p* at t1, and since he remembers everything, he should know that *p* at t2, too, but that's not the case. Therefore, Oscar does not have self-knowledge.

However, as I argued above in section 3.1 that no matter how long Oscar lives on twin earth and causally interacts with twater, provided he doesn't make any efforts to actively learn the concept TWATER, his thought contents would never switch. Having not made cognisant of the switch, Oscar does not suspect anything and consciously does not learn the twin concept TWATER. Since he never learned it, he only knows about the concept WATER and thus, I argue that Oscar's thought contents never change from p to p'. Therefore, it is not the case that when at t2 Oscar upon remembering his thought at t1, he remembers that *p'* and not that *p*. Thus, I reject premise 3 in Boghossian/Ludlow's memory argument and avoid the incompatibilist's conclusion.

Thus far, I argued that if Oscar does not know that he had been switched to twin earth, he would not consciously make efforts to learn the new concept, and thus, in effect, his thought contents would remain the same i.e. water thoughts, which does not lead to him lacking self-knowledge. I now consider a case where Oscar puts in the effort and acquires the new twin concept. Although, the domain of discussions in memory argument usually do not consider cases where Oscar becomes cognisant of the switch, John Gibbons mentions that loss of knowledge is possible *only* [emphasis mine] if Oscar becomes aware of the switch (Gibbons, 1996). Thus, it has become paramount to address such a case, too, and defend my claims.

Let us now suppose that Oscar was told that he had been transported to twin earth, say, at t*, where t* is some time in between t1 and t2 (imagine t1 as 2009, t2 as 2019 and then t* shall be, say, 2012)[6]. Upon this information, let us suppose that Oscar consciously learns about the new concept TWATER. As the only difference between water and twater is the molecular structure, the only way for Oscar to be able to discern them is to learn the molecular differences between the two. If concepts were to be thought of as mental representations (RTM)[7]—given the lack of consensus, I favour RTM just for the sake of

[6] As is customary of philosophical debates, especially when thought experiments are concerned, often little to no attention at all is provided on what kind of impact will we see in the real world.If such would happen, commonsense tell us that we would see Oscar revolting, not willing to cooperate, him slinging anti-government (or whoever responsible) slogans for his transportation without his consent, and so on. I would like to acknowledge this deficiency but wish to carry on. So, for the sake of argument, let us suppose that Oscar makes peace with his current situation and tries to adapt to the new environment.

[7] There are two other theses regarding the ontology of concepts: i) as the ability to discern one concept from the other (like ability to tell apart CAT from non-CATS) and ii) Fregean senses. It does not matter what conceptual ontology we subscribe to as it is not relevant to this thesis and does not have any effect on this subject matter.

explicature here—then both the concepts WATER and TWATER would include, in Oscar's mind, almost identical representations[8]. The only way through which Oscar can tell them apart is through inculcation in their representations of the idea (somehow) that their molecular structures vary, namely that one is made up of H_20 and that the other of XYZ.

If Oscar entertained a thought that water is wet at t1, and remembers it at t2, then having learnt that two mutually indistinguishable objects, namely that water and twater, exist, he would naturally wonder if what he had had was water thought or twater thought. But if he remembers that t1 was before his switch and that he learnt of twater at t*, he would be able to ascertain that he thought of water since he did not know of twater until t*.

Even if he has no clue that t1 was at a time before his switch (if we don't say when exactly his switch occurred) he would be still able to explicate that he may have thought of one of the two kinds of almost-indistinguishable objects and that he is unable to say which one he thought of at t1. Thus, in this case, too, it is farfetched to declare that Oscar does not know that *p* at t2; he knows that *p* or *p'* at t2, which is not enough to deny him self-knowledge (more on this below). Self-doubt does not exclude self-knowledge; only not knowing what one thought of does. Being confused about what one thought of is not sufficient ground to claim that Oscar has no self-knowledge.

In order to discuss this above case, we must look at the following two kinds of cases. Burge (2013) demarcates what he terms Disjunct Type cases and Amalgam type cases. The former is when two different concepts coexist in an agent's mind and one cannot be applied for the other; the agent may, however, mistakenly apply one for the other if they are of the water/twater type objects (i.e. nutually indistinguishable). The latter is when there are two concepts of which one is essentially the broadening of the other concept. For example, in arthritis/tharthritis case, the concept THARTHRITIS applies to inflammation of *both* the joints and the muscles, whereas the concept ARTHRITIS applies *only* to the inflammation of joints. Thus, if one were to use, say, twin Bert, who was slow-switched to earth, who has no idea of the earthian concept ARTHRITIS, the term

[8] Curiously, if an incompatibilist were to grant that, practically speaking, WATER and TWATER are the same since they both have the same functions (quenching, boiling, etc.), and that there is no point in distinguishing them two, they collapse under the weight of their own argument as, if one grants this, one cannot claim that Oscar does not know his own thought content, for in any case, he will have knowledge of either water or twater and that since they are both the same, there is no switch in Oscar's thought contents.

"tharthritis" to mean he has pain in his knees, it will be a valid usage[9]. Amalgam type cases are, thus, cases where one (twin) concept is essentially a *broadening* of the other concept. Having categorised these two cases, Burge argues that neither disjunctive or amalgam type cases make sufficient grounds to grant that the agent has no self-knowledge (Burge, 2013). Also, as a side note, Kevin Falvey takes a radical stance and argues that mental contents from memory are individuated by present environment, and grants that in order to have authority over our memorial thought contents it is enough for the agent to merely possesses a disjunctive concept,—in the terminology he uses,—namely, ZWATER (Falvey, 2003). Thus, disjunctive or amalgam type cases do not threaten our having authoritative self-knowledge.

But even without resorting to the radical stance that Falvey (2003) takes, I can still save my argument: if a critic were to add that in the debate over incompatibility, philosophers usually have something stronger in their mind (non-disjunctive or non-amalgamated type case) when they talk about first person authority and that my allowing disjunctive type cases weakens our notion of privileged access to our mental contents, then, to such a critic, I would answer that, yes, it usually is the case but my knowing that I thought of that-p-or-p' is more direct and authoritative than my knowing that you thought of that-q-or-q'. In this way, even in disjunctive/amalgam type cases, I have far more privileged access than I can have have of your mental contents. Thus, I think this objection does not threaten my claims here.

In this subsection, I attacked premise 3 of the memory argument due to Boghossian. I rejected premise 3 in all possible scenarios: when Oscar has no clue that he has undergone slow-switching, when Oscar was told of the switch but does not know when exactly he was switched, and when Oscar was told of the switch and exactly when. In all the scenarios, we can see that premise 3 does not hold up. This concludes my first strategy against the memory argument. In the next subsection, I forward a different strategy to attack, again, premise 3.

[9] Note that the amalgam type concepts cannot work under all conditions. For example, in this case, if twin Bert uses the term 'tharthritis' to refer to the inflammation of his muscles (and not joints), he will get corrected by the earthian doctor which would lead to narrowing of his concept THARTHRITIS. It is not known what would happen if twin Bert was not informed of the switch and happens to accidentally have such a "narrowing" of his twin concept. As this is not such a pressing issue and as this is not concerned to this thesis, I shall skip ruminating on this further.

3.2.2 RECONSTRUCTIVE MEMORY

There indeed is a huge chasm between philosophers of mind and cognitive sciences/psychology. Although brain studies originated as a field of subjecting to empirical observation what philosophers of mind have been positing for centuries, philosophers are often blind to the developments in these domains[10]. This subsection deals with one such 'missing out' by philosophers of developments in psychology that led them to argue in ways that do not agree with relevant empirical data about memory and remembrance. In the above section 3.2.1, I rejected premise 3 in Ludlow's formulation as a consequence of my arguments in section 3.1. In this subsection 3.2.2, I provide a different line of argument to reject, again, premise 3.

Attempts at undermining the memory argument usually challenge it on the grounds of entailment, on the grounds of how different memory as knowledge is, and so on (Bernecker, 2004; Brueckner, 1997; Falvey, 2003; Goldberg, 1997). I reject premise 3 on the grounds that the assumption made by incompatibilists on how memory works is different from how it actually works. To reach to that end, firstly, I argue that Ludlow's contention that contents of memory are individuated by current environment rests on a fallacious assumption; secondly, I provide an account of how memories actually work. Putting these two together, I find myself in an able position to reject premise 3 and thus, the memory argument does not lead to an incompatibilist conclusion. Let us look at my first argument now:

Ludlow, himself, an externalist argues that content externalism does not undermine self-knowledge in (Ludlow, 1995b). Although I, too, argue for the same ends, I find it implausible the way Ludlow argues: he posits that we defer to the linguistic community for the contents of all our mental contents (including memories), and holds that the contents are determined by current environment. He says, "...I am prepared to defer to the members of my linguistic community for the content of 'chicory' and this deference applies likewise to my memories." (Ludlow, 1995b, pp.158). On individuation of memorial contents, in his own words:

[10] There are, nonetheless, of late, philosophers of mind like Peter Carruthers who closely follow cognitive sciences/psychology and its developments and apply the learnings in their philosophical investigation (Carruthers, 2011).

> "...But notice that that recollection will have its contents fixed by current environmental conditions. As those environmental conditions shift, the content of my memory of this second-order mental episode will shift (just as the contents of memories of first order thoughts shift)." (Ludlow, 1995b, pp.159).

That is, Ludlow endorses presentist externalism when it comes to memorial contents; he holds that contents of memories, too, like occurrent thoughts, are individuated by current environmental conditions. The other two kinds of externalism that are available are pastist and futurist externalism which holds that contents of memories are individuated, respectively, by past and future environmental conditions. Of these stances, futurist externalism is not endorsed by many philosophers with few exceptions such as Henry Jackman (1999). However, even intuitively, I think that pastist externalism makes much more sense than presentist externalism. Does it not feel counterintuitive to think that the current environment has the ability to tamper with things we actually remember, the contents of which must have been, provided we do not forget anything, already fixed in our mind?[11] Occurrent thoughts may very well be individuated by current environment but contents of memory can only be individuated by past environment, as memorial contents are not about what is the case now, but is about *what was the case* back then. It could be the case that both past and present environments have remained unchanged but that is not to say that the current environment individuates memorial contents; only the past environment can.

Bernecker succinctly puts the notion that individuation of the contents of memory are dependent not on current environment but on the original environment in which the memory was formed as follows:

> "The memory claim is true if and only if it is an accurate representation of some past states of affairs. Given temporal externalism, however, the concepts used in memories are determined by the present environment. But if the concepts employed in memories refer to present affairs so do the truth conditions of memories. And if the truth conditions of memories refer to present affairs, the memory claim in question would turn out false if there were no longer a tree in S's garden. But this

[11] When I write "fixed in the mind," it might seem that I suggest a form of content being stored in brain's "memory bags." But I use this only as a manner of speaking since I'm not advocating this view. The argument below deals precisely with this.

> is surely absurd. [...] Thus I conclude that the only viable form of externalism about memory says that memory contents are determined by past environmental conditions." (Bernecker, 2004, pp.611-612).

Thus, having seen reasons to reject presentist externalism and embrace pastist externalism, let us move on to see how memory actually works, which is pretty much at loggerheads with what philosophers in this debate considered to be the case:

It was long held that our memories are stored in different parts of the brain and that every time we recollect,—that is when we attempt to decode—we reactivate the same neural activity—the same, as it were, mental pathway—in the brain that initially led to the encoding of the contents of our memories. This standard folk view sees memories as things that are essentially stored in "memory bags," and that we revisit the same neural pathways to retrieve the content every time we attempt to recollect memories. But this view has been debunked since the 1930s but still remains among the common folk talk. The present, widely-accepted scientific theory of memory is that memorial recollection is reconstructive (Bartlett, 1932). That is, every time we recollect, we essentially reconstruct the memories using the stimuli we already have. We put together the stimuli in as exact a manner as possible to recreate our memories. Despite the fact that there have been few objections, by now the view that memory is reconstructive is pretty much the unanimously accepted view in the cognitive sciences and psychology ("Memory and Learning," 2002; Wagoner, 2017). Indeed, this phenomenon is even more pronouncedly observable in socio-cultural settings (Wagoner, 2012). In Sir Frederic Bartlett's own words:

> "Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction or construction, built out of the relation of our attitude towards a whole active mass of organised past reactions or experience, and to a little outstanding detail which commonly appears in image or in language form. It is thus hardly ever really exact, even in the most rudimentary cases of rote recapitulation, and it is not at all important that be should be so." (Bartlett, 1932, pp.213-214).

Granted that recollection of memories are nothing but reconstructions, there remains the question of how reconstruction is done. The answer is that reconstruction is done using the stimuli from the past environment. I argued above from a philosophical

perspective that contents of memories are individuated by past, and not present, environment. It is also supported by empirical research which reassures us that the stimuli that reconstruction uses upon recall are stimuli from the past, provided one does not forget relevant details (Bartlett, 1968); if one does forget parts, or all, of the relevant stimuli from the past, one's reconstruction is done using various other fitting stimuli to complete a picture which leads to the creation of false memories. Thus, misremembering, false memories, etc., are results of either misapplying the 'assembling tools' during reconstruction or of missing some stimuli. Thus, the theory of reconstruction can accommodate both correct and false memories making it a wholesome theory.

Putting these two together,—reconstructive theory of memory and pastist externalism—it becomes apparent that when Oscar tries to remember what he thought of at t1, as the contents of his memory are individuated by the stimuli from the past environment, and, as he forgot nothing as granted by premise 2 (all the stimuli his mind would use in reconstruction is guaranteed to remain intact), he will have an accurate memory of what he thought of at t1. In other words, aided by accuracy in memorial reconstruction and content individuation by past environment, Oscar can have an absolute recollection that p—the thought he had at t1—also at t2, thereby rebutting premise 3.

My reliance on the theory of memory as reconstructive might come off as a surprise to many a critic, for the preservationist view that philosophers in this debate favoured (or, took for granted), at least, had, at its core, the property of "freezing up" (fixing) past contents in one's brain, whereas reconstructivism is notorious for destructing this happy picture of human memory. At the outset, my invoking reconstructivism, as it were, strengthens the camp that proponents of memory argument reside in, but we can see that my arguments succeed even in this reconstructivist view which does not, at its core, have this property of "fixing" past contents. Indeed, one of the central tenets of reconstructivist theory of memory is that the past contents need not be fixed in one's mind. But, as I argued above, even this does not bother the compatibilist picture that I paint, given constraints such as not forgetting any relevant stimuli. So, to summarise my memory reconstruction argument: I pointed out that individuation of contents of memory depends on the past, and not present, environment; and, having done that, I argued that since recollection of memories is reconstructive, Oscar will fail to have water thoughts only if he forgets the relevant stimuli—such as the concept WATER. But premise 2 of the memory

argument guarantees that Oscar will not forget any relevant stimuli, and therefore, Oscar will be able to remember his water thoughts that he had at t1 at t2, too. Thus, premise 3 that Oscar does not know that p is not true. Hence, we can abstain from reaching an incompatibilist conclusion.

Finally, before concluding this chapter, I summarise below the central claims and structure of the memory argument, followed by my opposing arguments:

1. If Oscar does not forget anything relevant, then what he knows at t1, he knows at t2.
2. Oscar forgot nothing relevant.
3. By modus ponens from 1 and 2: if Oscar knows that p at t1, then he knows that p at t2.
4. If we have authoritative self-knowledge, then at t1, Oscar knows that Oscar thinks that p at t1.
5. 3 and 4 logically entail that if we have authoritative self-knowledge, then at t2, Oscar knows that he thought that p at t1.
6. If content externalism is true, then Oscar does not know that p at t2.
7. If Oscar does not know that p at t2, then he does not know at t2 that he thought that p at t1.
8. 6 and 7 logically entail that if content externalism is true, then at t2, Oscar does not know that he thought that p at t1.
9. 5 and 8 logically entails that it is not the case that externalism is true and that we have self-knowledge (or, in other words, that they are incompatible).

In my argument against this formulation, I reject premise 6 that if externalism is true, then Oscar does not know that p at t2. To better illustrate, let us replace p with "water". I reject premise 6 with the following argumentation (we assume that content externalism is true for this argument and so I start with treating the consequent alone: that Oscar does not know that p at t2):

1. Through my arguments in section 3.1, Oscar, in the case considered, does not acquire the concept TWATER at t2.
2. Thus, his thought contents at both t1 and t2 express only WATER concept.

3. If he only had the concept WATER at both t1 and t2, his memorial contents cannot have TWATER concept in them.
4. Thus, at t2, Oscar, upon recollection, can express the concept WATER (thus, knows that p at t2)—this concludes the first part of my arguments in 3.2.1.
5. Oscar will be able to learn and acquire TWATER concept after he was made cognisant of the switch.
6. If he acquires TWATER concept sometime between t1 and t2, then he will be able to refer to twater when he uses "twater" at t2.
7. In the case where he knows that t1 was before the switch, at t2, he will be able to know that he had water thought at t1.
8. In the case where he does not know when t1 happened, at t2, he will be able to know that he either had water or twater thought at t1.
    a. Disjunctive type cases are not enough grounds to grant that Oscar does not know his own thought contents.
9. By 7 and 8, whether or not Oscar knows when t1 occurred, at t2 Oscar can know that he had water thoughts (or a disjunctive thought with water thought as one of the components) at t1.
10. Thus, at t2, Oscar, upon recollection, can express the concept WATER (thus, knows that p at t2)—this concludes the second part of my argument in 3.2.1.
11. If memory is reconstructive, then Oscar's thought contents are individuated by his past environment.
12. Memory is reconstructive.
13. Thus, Oscar's memorial contents depend on his past environment.
    a. At t1, Oscar's environment had water.
    b. Memorial reconstruction would use this stimulus, namely water, to reconstruct his thoughts.
14. If Oscar forgets nothing relevant, then his memorial recollection will reproduce exactly what he thought of at t1, namely water thoughts.
15. Oscar forgot nothing relevant.
16. Thus, at t2, Oscar, upon recollection, can express the concept WATER (thus, knows that p at t2)—this is my argument in 3.2.2.
17. Thus, through 4, 10 and 17, in all the cases considered, Oscar knows that p at t2.

Thus, I reject premise 6 in my above summary and avoid being led to the incompatibilist conclusion. With this, I complete my arguments against the plausibility of slow-switching cases in establishing incompatibility between content externalism and authoritative self-knowledge. In this way, I point out that premise 3 of the memory argument can be rejected as Oscar's memorial thought contents do not change unbeknownst to him upon recollection. This is true of both the cases where Oscar knows of the switch (and, thus, learns the concept TWATER) and does not know of the switch as I argued above. Therefore, the memory argument does not succeed in demonstrating that content externalism is incompatible with authoritative self-knowledge.

Before concluding the thesis in chapter 5, I address four potential objections in the next chapter.

4. REPLIES TO POSSIBLE OBJECTIONS

This chapter examines some of the possible objections that my arguments above may receive, and provides my responses to them.

4.1 ON DEFERENCE

One can object to my argument that thought contents can change unknown to the agent through the mechanism of deference. We can defer to experts in the linguistic community and have our thought contents individuated; have the community fix the contents of our thoughts unknown to us. However, I think that deference results in the agent thinking that he is talking about something but the community, with the agent's deference, interpreting him differently. The agent will acquire the new concept only if the agent can himself distinguish between the two concepts. Until then the agent might defer to the community but still may fail to possess the new concept.

To see how this would work, let us move away from the problematic water/twater example and take Burge's arthritis example for socio-linguistic externalism[12]. Consider the case of twin Bert who lives on twin earth and whose linguistic community uses tharthritis to refer to ailments of both the joints and the muscles. Suppose that twin Bert was switched to earth without his knowledge and lives on earth for a long time. Let him causally interact with arthritis patients on earth who may use various linguistic expressions with the word "arthritis." Twin Bert, of course, does not have any idea that by "arthritis," earthians express a different concept and that he, as a twin-earthian, possesses a different concept for the same word. During his usage, he defers to the experts in earthian community. Since the twin-earthian concept encompasses both the ailments of thighs and joints, when twin Bert, who develops an inflammation on his knees, use that word, earthians do not suspect anything and interpret him as though he meant arthritis; and, as tharthritis applies to this case, too, Twin Bert has no reason to suspect the earthian usage, and assumes that they use

[12] As is pointed out by countless subsequent literature that Putnam inspired, the usage of water as an example has many unwanted side effects such as human body being made up of 70% water (H_2O), or that natural kinds being rigid designators, and so on.

the same concept that he uses[13]. He defers to the experts and if the deference were to bestow him with the earthian concept he would have known that he cannot use the word for inflammation of his thighs. But he still retains his twin earthian concept THARTHRITIS and always expresses on earth, too. Thus, despite deferring, he fails to have his thought contents changed. Even if he, by pure chance or wild guess, does become suspicious, he would investigate and would end up learning of the (new) earthian concept. This case, too, fails to threaten my argument as his thought contents changed now with his awareness, with his conscious efforts. Now let us suppose that he develops an inflammation in his thighs. Having learnt on twin earth that tharthritis can occur (also) on the thigh, twin Bert goes to an earthian doctor and reports (by which he defers to the experts), "I have tharthritis on my thighs." The earthian doctor on earth would correct twin Bert that arthritis is an ailment that applies only to inflammation of joints and not thighs. Then twin Bert would have come to learn through active cognitive participation the new concept ARTHRITIS. In this case, suppose post his visit to doctor he was told that he had undergone slow-switching, he would realise that there are two different concepts and would learn of factors that distinguish arthritis and tharthritis. The advantage this example has is that it illustrates that twin Bert's causal interaction with earthian object—namely, arthritis, in this case—is not enough to bring about a change in his concepts. Thus, in all the cases whether or not he was made aware of the slow-switching, Oscar can continue to defer to experts but it is still not enough for him to acquire a twin concept.

By that, I do not deny that one can learn concepts through deference. Obviously, if one comes across a new concept, in deferring to experts, one learns of its usage, its applicability, its truth conditions, and so on, which are enough to grant that this agent possesses this concept this agent is not previously exposed to. But deference in the slow-switching cases, where there are two mutually indistinguishable concepts available, does not have any effect on the agent acquiring the twin concept. That is, Doscar from Dry

[13] The case, of course, would be different if we switch earthian Bert to twin earth. He would, at the very first instance, when a twin-earthian shows his puffed-up thigh and says, "I have tharthritis," would attempt to correct this twin-earthian person, and, that would, much to his surprise or dismay, lead him to discover that he had learnt it all wrong. Post learning this, he may even abandon his previous concept (this is assuming that he has no clue that he had been switched; if he knows of the switch, he, would, of course, attempt to acquire tharthritis concept and retain both the concepts). His abandoning the earthian concept in the former case is a different matter altogether (he knowingly dropped it) and, for the purposes of the present argument, is inessential.

Earth (an earth that does not have any water), when exposed to water (or, slow-switched to earth), can, through deference, causal interaction, etc., come to possess the concept WATER but when, again, slow-switched to twin earth, will not be able to acquire TWATER unless and until he learns of the molecular differences between water and twater. Doscar's case, you will notice, is akin to that of a newborn baby which necessarily need not consciously learn of the concepts WATER, SOFA, ARTHRITIS, etc., when she first encounters a concept, but if slow-switched to another relevant alternative environment, she will fail to acquire the twin concept unless she becomes capable of distinguishing between the two concepts.

4.2 NATURAL KINDS AS PRIMITIVE CONCEPTS?

A persistent critic can point out the possibility of natural kinds being a primitive concept in our mind's conceptual schema. If natural kinds turn out to be one of those unlearnable, innate, primitive concepts, then Oscar can, in the presence of twater, automatically come to acquire the concept TWATER—why, he can acquire it even when there is a poverty of this stimulus! To such a critic, I would say that, yes, it could be the case, but even in that case, as I already did point out in section 3.1, (i) natural kinds are not primitive concepts, and (ii) even if we are wrong about that, primitive concepts must nonetheless be learned (Margolis, 1998; Laurence and Margolis, 2011). Also, just to clarify: learning could be semi-/unconscious, too. This point was raised to rebuff the notion that there are certain concepts that cannot be learned.

Does this mean that all concepts are to be learned consciously by the agent? Am I saying that this is how infants learn? No, the need for conscious learning is applicable only to concepts that are indistinguishable superficially. Obviously children, or even adults, when they acquire a new concept don't make conscious cognitive efforts to learn the concept. They learn to grab a concept just by interacting with it causally. Say, when a child first learns what a bolt is she need not know the microstructure, how it is used, and so on. But upon encountering two different types of bolts that look exactly similar (say, in engineering school) and are made up of the same material, she should learn to distinguish that the two bolts are made up of exactly the same molecules but differ only in how they were heat-treated that gives rise to each of them slightly different physical properties. That

is, it is possible for a screw and a bolt to be both be made up of a same grade of medium carbon steel, subjected to same types of heat-treatment to harden it, and so on, so that both end up having the same physical and chemical properties. The only difference between these two is their use—bolts are used in joints that are unthreaded and always take up a nut, whereas screws are used in joints that are threaded and do not take up a nut. It is not possible to know that SCREW and BOLT are different concepts without knowing the conventions associated with it. No matter how long a person spends on a twin environment (earth, engineering facility, etc.), causally interacting, in such almost-indistinguishable cases—like those the entire slow-switching arguments are based on—change in their mental contents will not take place.

I, thus, only argue against automatic content switch as was supposed to happen by incompatibilists who forward the slow-switching thought experiment. It is possible for an agent who has never been acquainted to a concept of a certain kind—say, a new social convention—to acquire it just through causal interaction. If this agent were to get exposed to something indistinguishable from this particular concept, then the agent should, as I suggest, consciously learn to distinguish the two. Until then the agent will keep using linguistic expressions and thoughts that expresses, and refers to, the older concept that the agent is already acquainted with.

4.3 ON THE PROSPECT OF MISCOMMUNICATION

If Oscar, in his interactions with twin-earthians, means water whereas the twin-earthians interpret him as meaning twater, doesn't it indicate that there is a miscommunication here? Doesn't that mean that every time an agent switches their environment, unless the agent learns of the new concept (or defers to the community), there is a miscommunication? We often change our environment (such as Ludlow's examples of Biff and the meaning of "pragmatism" in two different communities in (Ludlow, 1995a)) and, if I were to latch on to my claims, it might suggest that there is always a constant miscommunication when we speak to others.

To this objection, I bite the bullet and grant that a harmless miscommunication ensues. Even in normal day-to-day conversations of very ordinary concepts, it is highly implausible to suppose that one person understands the other person perfectly. No one is

able to communicate exactly what one thought of, what one meant to convey, and yet the recipient understands much of what the communicator intended to express and acts in accordance to it. There is always a miscommunication in every single interaction we have with others but yet manage to get by without any problems. This cannot be downplayed as the problem of mediation of one's thoughts through language is considered one of the central problems in ordinary language philosophy. I, thus, appeal to such commonsensical transactions where people are totally fine using mutually indistinguishable concepts at no cost to themselves. Two concepts C1 and C2 may coexist in an agent's mind unless severe repercussions/misunderstandings take place. If the latter occurs, the agent will be forced to distinguish the two concepts, and the agent may do it with the aid of some conceptual learning framework (such as molecular differences in our water/twater scenario). By "harmless," I mean that it will not give rise to any severe repercussions. If there be any severe repercussion—such as being grossly misunderstood—that would lead to the agent realising that the nature of the concept they possess might be wrong/outdated, and they then would change/update it accordingly (or will be unwilling to change, but the point here is that they will become aware that there is a difference). Thus, to repeat, I grant that there is a miscommunication only in the theoretical sense of the word; for all the practical purposes, this does not matter and that the agent will be able to succeed in their transactions with the external world.

This might appear to be an extremist view of communications, as one can argue that communications need not be perfect in order to be understood. But if one is willing to permit the latter why not the former?

4.4 AGAINST ABSOLUTE RECOLLECTION

Now one might object that it is not possible to forget nothing and recollect absolutely. We have to, in line with our fantastical thought experiment, assume that Oscar's memorial reconstruction will be perfect. I mention this only in order to provide the memory argument a charitable reading. But if the critic were to press even further, I can argue that the possibility of absolute recollection is not as far-fetched as it may sound. Indeed, Bartlett himself is not against—what I term—absolute recollection. He writes:

> "I did not say, I think I did not imply, that literal retrieval is impossible, but I did imply that it requires special constricting conditions. I specified certain social conditions in which this seems particularly likely to occur. [...] Nothing that I wrote was intended to deny the possibility of this, and if any of the statements in the theoretical parts of the book seem to imply such denial they must have been badly phrased." (Bartlett, 1968, section 3).

In special constricting situations, absolute recollection is possible and there were even famous cases of people being able to recount every event in accurate detail about every single day of their lives. This is, of course, not germane to the current domain of discussion but is only meant to showcase that what a critic might take as preposterous is actually possible, and, that, is very much consistent with our thought experiment.

Leaving aside the fantastical thought experiment, we are faced with one more question: what would the actual role be that the environment can play? It will only be along the following lines. We would, I maintain, still have the same content in our memories, however diluted we may feel the associated propositional attitudes be—indeed, this is the case with our memories: when we remember certain mental episode, we certainly retain the same propositional attitude but in time, it may get diluted (or, get enhanced, in certain rare cases)—as is the case with traumatic memories, happy, sad thoughts, and so on. Dilution occurs usually when you forget all the nitty-gritty details of the memory (with premise 2 ensuring us that dilution will not happen to Oscar in our thought experiment). That is not to say that, provided the agent forgets none of the relevant stimuli from the past, the external environment will not affect an agent's memory at all, but only that the external environment can dilute (or, enhance, in rare cases) the vividness of an agent's memory, but it does not, and cannot, switch the inner content. Changes in contents of the agent's memory will take place only in cases where the agent forgets parts of the past stimuli and replaces the missing parts with available parts, thereby leading to misremembering, formation and recollection of false memories, memorial delusions, etc. But, coupled with my argument that contents don't change, even this case does not lead us to the incompatibilist conclusion.

So far in this chapter, I addressed four potential objections to my thesis and attempted to defend my stance. I take it that my defence is successful. Thus I conclude that

the arguments advanced through slow-switching cases do not succeed in establishing the incompatibility thesis.

5. CONCLUSION

In this thesis, I argued that both the variants of the slow-switching arguments that aim to establish the incompatibility thesis—that content externalism is incompatible with authoritative self-knowledge—fail in its aim owing to two reasons: the first one being a rejection of the assumption that an agent's concepts can automatically change through causal interaction with their surroundings, and the second one being the nature of recollection of memories not permitting the changes in memorial contents of the kind claimed by incompatibilists to occur. Thus, it can be concluded that slow switching arguments do not succeed in establishing the incompatibility of content externalism and authoritative self-knowledge. Having said that, an obvious question now remains: if authoritative self-knowledge and content externalism are compatible, what is the relationship between the two? What kind of influence does the external environment exert on our authoritative self-knowledge, or how does it inform us, if at all, of our own thoughts? As this is beyond the scope of this thesis, I shall not take up this now. This could be the subject of another paper.

6. BIBLIOGRAPHY

Bartlett, F. C. "Remembering: A Study in Experimental and Social Psychology." (1932).

Bartlett, Frederic Charles. "Notes on remembering." *Sir Frederic Bartlett Archive* (1968).

Ben-Yami, Hanoch. "The semantics of kind terms." *Philosophical Studies* 102.2 (2001): 155-184.

Bernecker, Sven. "Memory and externalism."Philosophy and Phenomenological Research 69.3 (2004): 605-632.

Boghossian, Paul A. "Content and self-knowledge." Philosophical Topics 17.1 (1989): 5-26.

Boghossian, Paul A. "What the externalist can know a priori." Proceedings of the Aristotelian

Society. Vol. 97. Aristotelian Society, Wiley, 1997: 161-175.

Brueckner, Anthony."What an anti-individualist knows a priori."Analysis 52.2 (1992): 111-118.

Brueckner, Anthony. "Externalism and memory." Pacific Philosophical Quarterly 78.1 (1997): 1-12.

Burge, Tyler. "Individualism and the Mental." Midwest studies in philosophy 4.1 (1979): 73-121.

Burge, Tyler. "Individualism and self-knowledge." The Journal of Philosophy 85.11 (1988): 649-663.

Burge, Tyler. "Wherein is language social?." (1989) in "Propositional attitudes: The role of content in logic, language, and mind." ed. by Anthony and Owens (1990).

Burge, Tyler. "Memory and Self-Knowledge" in Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays, Volume 3.: Oxford University Press, May 23, 2013

Carter, J. Adam, et al. "Varieties of externalism." Philosophical Issues 24.1 (2014): 63-109.

Carruthers, Peter. The opacity of mind: An integrative theory of self-knowledge. OUP Oxford, 2011.

Falvey, Kevin (2003). Memory and knowledge of content. In Susana Nuccetelli (ed.), New Essays on Semantic Externalism and Self-Knowledge. MIT Press.

Gibbons, John. "Externalism and knowledge of content." *The Philosophical Review* 105.3 (1996): 287-310.

Goldberg, Sanford C. "Self-ascription, Self-knowledge, and the Memory Argument." Analysis 57.3 (1997): 211-219.

Falvey, K. & J. Owens, 1994, "Externalism, Self-Knowledge, and Skepticism," Philosophical Review, 103: 107–37.

Jackman, Henry. "We live forwards but understand backwards: Linguistic practices and future behavior." Pacific Philosophical Quarterly 80.2 (1999): 157-177.

Laurence, Stephen, and Eric Margolis. "The poverty of the stimulus argument." The British Journal for the Philosophy of Science 52.2 (2001): 217-276

Leslie, Alan M. "How to acquire a 'representational theory of mind'." Metarepresentations: A multidisciplinary perspective (2000): 197-223.

Ludlow, Peter. "Externalism, self-knowledge, and the prevalence of slow switching." Analysis 55.1 (1995a): 45-49.

Ludlow, Peter. "Social externalism, self-knowledge, and memory." Analysis 55.3 (1995b): 157-159.

Ludlow, Peter. "First Person Authority and Memory." Interpretations and Causes. Springer,

Dordrecht, 1999. 159-170.

Margolis, Eric. "How to acquire a concept." Mind & Language 13.3 (1998): 347-369.

Margolis, Eric, and Stephen Laurence. "Learning matters: The role of learning in concept acquisition." *Mind & Language* 26.5 (2011): 507-539.

"Memory And Learning". *Thebrain.Mcgill.Ca*, 2002, http://thebrain.mcgill.ca/flash/i/i_07/i_07_p/i_07_p_tra/i_07_p_tra.html. Accessed 3 Apr 2019.

Morvarid, Mahmoud. "The epistemological bases of the slow switching argument." European Journal of Philosophy 23.1 (2015): 17-38.

Parent, T. "Externalism and "knowing what" one thinks." *Synthese* 192.5 (2015): 1337-1350.

Pryor, James. "What's Wrong with McKinsey-style Reasoning?." *Internalism and Externalism in Semantics and Epistemology* (2007): 177-200.

Putnam, Hilary. "Meaning and reference." The journal of philosophy 70.19 (1974): 699-711.

Smith, Andrew F. "Semantic externalism, authoritative self-knowledge, and adaptation to slow switching." Acta Analytica18.30-31 (2003): 71-87.

Tobia, Kevin P., George Newman, and Joshua Knobe. "Water is and is not H2O." (2017)

Vahid, Hamid. "Externalism, Slow Switching and Privileged Self‐Knowledge 1." Philosophy and Phenomenological Research 66.2 (2003): 370-388

Wagoner, Brady. "Culture in constructive remembering." *Oxford handbook of culture and psychology* (2012): 1034-1054.

Wagoner, Brady. "What makes memory constructive? A study in the serial reproduction of Bartlett's experiments." *Culture & Psychology* 23.2 (2017): 186-207.

Warfield, Ted A. "Privileged self-knowledge and externalism are compatible." Analysis 52.4 (1992): 232-237.

Warfield, Ted A. "Externalism, privileged self‐knowledge, and the irrelevance of slow switching." Analysis 57.4 (1997): 282-284

ABSTRACT

ON THE IMPLAUSIBILITY OF SLOW-SWITCHING ARGUMENTS IN ESTABLISHING INCOMPATIBILISM

Philosophers who argue that content externalism is incompatible with authoritative self-knowledge usually employ one of the two arguments namely the slow-switching argument and the reductio ad absurdum. Of these I focus on only the former which in itself has two variants namely the content-switch (main argument) and the memory argument (a variant). I argue against both the variants thereby denying that slow-switching arguments succeed in establishing the incompatibility thesis.

It is long held that if a slow-switched agent (Oscar) were to stay long enough on twin earth, his thought contents change unbeknownst to him. And it was reasoned that, since Oscar is unaware of the changes in his mental contents and cannot spot when the changes occurred, he does not have access to his own thought contents at all times, which thereby leads to the conclusion that authoritative self-knowledge is incompatible with externalism. In this thesis, I argue that, in cases like these, mental contents do not change unknown to Oscar. I appeal to theories of concept acquisition to achieve this end. This forms my attack on the main argument. And, as against the memory argument, I use two strategies the first one of which is an extension of the previous argument applied to this case; and the second strategy is to argue that memorial recollection depends on the past, and not the present, environment and, if Oscar does not forget any relevant past stimuli, his memorial contents upon recollection will not change. Having thus argued against both the variants of the slow-switching arguments, I conclude that slow-switching arguments do not succeed in establishing the incompatibility thesis.

I, Arunkumar Rajavel,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

On the Implausibility of Slow-Switching Arguments in Establishing Incompatibility Thesis supervised by Prof Bryan Frances and Prof Juhani Yli-Vakkuri

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Arunkumar Rajavel
Tartu, 15.05.2019