Externalist Internalism

Essays on Hybrid Theories in the Philosophy of Mind and Epistemology

Milena Bartholain

New College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

74.830 words

January 2023

To my parents Stephanie and Lutz

Externalist Internalism

—

Essays on Hybrid Theories in the
Philosophy of Mind and Epistemology

Milena Bartholain
New College, University of Oxford
A thesis submitted for the degree of Doctor of Philosophy
January 2023

Abstract

This thesis collects four chapters connected by the theme of hybridity across theories in the philosophy of mind and epistemology. More precisely, this thesis assesses the consistency of hybrid epistemic internalism, a view that combines externalism about mental states with epistemic internalism.

I begin in **chapter 1** by motivating both views. I distinguish between two variants of epistemic internalism, access internalism and mentalism, that are motivated by two considerations, the access motivation and the Equal Justification Thesis. I defend a non-standard version of access internalism that focusses on the permissive notion of "*ready access*". I identify two challenges to the consistency of hybrid epistemic internalism: the access problem and the equal justification problem.

In **chapter 2**, I turn to the equal justification problem. State externalism undercuts the Equal Justification Thesis which states that individuals in the bad case have justification to believe the same things as their internal duplicates in the good case. I argue that a weak variant of the thesis, the Indiscriminable Justification Thesis, solves the equal justification problem if certain results of chapter 3 and 4 are established.

In **chapter 3**, I discuss the access problem in the form of the discrimination argument against ready access to one's external states. I defend the discrimination argument against a compatibilist strategy that argues either that the discrimination argument is not sound, or that its application conditions are not relevant to us. I conclude that state externalism leads to an access problem.

In **chapter 4**, I defend a novel epistemic definition of the internalism/ externalism debate about mind, the *categorical epistemic definition*. According to this view, state internalists and externalists disagree about the possibility of "*mental state switching*". The view implies that state externalism leads to an access problem by definition. Further, by emphasising the difference between an epistemic definition of the debate and metaphysically committing accounts within the debate, the proposed view avoids common objections.

Word count: 74.830

Thesis supervisor: Timothy Williamson
Title: Wykeham Professor of Logic

Thesis supervisor: Bernhard Salow
Title: Associate Professor of Philosophy, Tutorial Fellow at Magdalen College

Preface

This thesis isn't the continuation of long-burning passion but it is fair to say that it almost turned into one. During my BPhil, I took a seminar with Bill Child and Anil Gomes on the philosophy of mind where we read "What is Externalism?" by Katalin Farkas. I liked the big-picture challenge that she formulates in this paper against extant conceptions of the internalism/ externalism about mind. But I didn't know that I would find it so stimulating a couple of years later. Internalism/externalism in epistemology, however, has accompanied me ever since my undergraduate years and throughout my BPhil thesis on the KK-principle. This DPhil thesis was an opportunity to combine two debates on the fundamental relationship between us and the environment that we find ourselves in. Maybe unexpectedly, I enjoyed working on this thesis more the further it got. I properly understood debates I thought I knew quite well. It is responsible for recovering quite a bit of the joy of doing philosophy, and actually discovering new joys, after a longish period of frustration with it.

It is a privilege to think about something so fundamental as the nature of our mental states and, all the more, of those of the people on Twin Earth. I am thankful for the financial support I received to work on this thesis from the AHRC. New College supported me through the pandemic in the form of allowances provided by Old Members of the college on multiple occasions.

This thesis would not be what it is without the advice and the insights of my supervisors.

I am extremely grateful to Timothy Williamson who has not only supervised this DPhil thesis, but also my BPhil thesis. In both degrees, he has seen me through motivational lows and made sure that the quality of the work does not suffer from it (at least not in the final product). He didn't grow tired to highlight the complexities of issues when I had overlooked them. I believe that the spirit he taught me of holding oneself to higher standards will stay with me for whatever future task I will take on.

Bernhard Salow's comments and insights worked perfectly in combination with Tim Williamson's. Bernhard Salow gave as much invaluable advice about overarching connections and storylines as he was interested in running with ideas. He often raised issues and suggested solutions in one breath.

At the beginning of the phase when I (finally) knew which question I was going to focus on, I had a number of excessively long and fun meetings with Mike Martin. There were very encouraging to me at the time because Mike Martin was excited for me to think about such fundamental issues.

My time in Oxford was extremely important to me philosophically and privately, not least because I found a second home of heart. I made many very good friends among philosophers in Ben Brast-McKie, Simon-Pierre Chevarie-Cossette, Sam Clarke, Luke Davies, Christina Dietz, Alice Evatt, AJ Gilbert, Philipp Kurbel, Annina Loets, Chiara Martini, Sybilla Pereira and Weng Kin San. Chris Fowles, Matt Hewson and James Kirkpatrick are not only great friends but also gave very helpful feedback on material of this thesis.

I thank my partner Michael for making sure that my life besides the thesis was a great source of excitement without which I could not have put in the many hours of desk work that philosophy requires.

I dedicate this thesis to my parents because I would simply not have been able to write it without their generous support in all aspects of life. Even when I started writing about "water" that is XYZ, they were interested in the details of my work. They have been outstandingly supportive.

Contents

Chapter 1

1 - Hybrid Epistemic Internalism
Combining Externalism about Mind and Epistemic Internalism

1.1 Introduction: Hybrid Epistemic Internalism

Internalism, in most general terms, is the thesis that certain properties of philosophical relevance supervene on what is internal to an individual.[1] Externalism, in most general terms, is the denial of internalism: certain properties of philosophical relevance do not exclusively supervene on what is internal to an individual. Externalism further allows for external factors to determine whether the property of philosophical relevance is instantiated.

There have been two distinct debates making use of the labels "internalism" and "externalism": one in the philosophy of mind that is concerned with the fundamental nature of our thoughts; one in epistemology that is concerned with the nature of epistemic justification.[2] In the philosophy of mind, state internalists believe that mental states supervene on what is internal to an individual; state externalists deny that mental states supervene exclusively on what is internal to an individual and allow for external factors to determine the nature of mental states. In epistemology, epistemic internalists believe that whether an individual's belief that *p* at a time is justified supervenes on what is internal to that individual at that time; epistemic externalists deny epistemic internalism and allow for external factors to partly determine the justification of the belief that *p* at that time.[3]

As we will see, even though both debates use the same labels "internalism" and "externalism", the ways in which the internal has been traditionally understood in the two debates are not obviously related. The internal/external distinction in the philosophy of mind is understood *metaphysically*: roughly, an individual's internal properties are those properties that for their existence do not imply the existence of something

---

[1] One family of properties A supervenes on another B if and only if two objects cannot differ in their A-properties without differing in their B-properties (Kim 2002).

[2] In fact, there are more internalism/externalism debates. There is a second, less well-known internalism/externalism debate in epistemology between evidential internalism/externalism. I will come back to that in §1.1.2. There is also a internalism/externalism debate in metaethics that will not matter in the following.

[3] Belief is just one example of a doxastic attitude that individuals may justifiably take towards a proposition; others are disbelief and suspension of judgment. I will focus on beliefs for the sake of brevity.

apart from the individual.[4] In epistemology, the internal/external distinction is mainly understood *epistemically*: roughly, an individual's internal properties are those properties that they are in a position to know in a relevant way. If the internal/external distinction is understood differently in the two debates, then it would seem likely that an internalist position in one debate is not in opposition to an externalist position in the other debate, and *vice versa*.

This is interesting for the following reason. The correct theory of mental states would seem to matter to the correct theory of justification insofar as the doxastic states such as beliefs, disbeliefs and suspensions of judgment that get justified and that do the justifying are mental states. In the philosophy of mind, state externalism has become an increasingly popular position over the last forty plus years. Inconsistency with state externalism would be a severe cost for any justification theory. But if "internalism" and "externalism" are defined differently in both debates, it would seem that no such inconsistency would seem to be threatening.

Let us call a theory "hybrid" if it combines an internalist position in one of the debates with an externalist position in the other debate. In particular, let us call a theory "hybrid epistemic internalism" if it combines an externalist view of mental states with an internalist view of justification. This thesis is dedicated to assessing the consistency of hybrid epistemic internalist views.

In the rest of this introduction, I will proceed as follows. I will first introduce internalism and externalism about mind (**§1.1.1**). Then, I will turn to internalism and externalism about justification (**§1.1.2**). Here the situation is more complex because the traditional form of epistemic internalism, access internalism, conceives of the internal/external distinction in epistemic terms. The newer form of epistemic internalism, mentalism, however conceives of the internal/external distinction in metaphysical terms, just as it is done in the debate about mind. Epistemic externalist positions will vary accordingly. In

[4] Other labels that have been used to capture a metaphysical distinction along these lines are "intrinsic" as opposed to "extrinsic" properties, and "non-relational" as opposed to "relational" properties (see, e.g., Farkas 2003; Brown 2004; Gertler 2012).

the last section, I will show that hybrid epistemic internalism is the only hybrid view that has sparked a genuine interest in hybridity (**§1.1.3**).

1.1.1 Internalism/externalism about mind

When I say that internalists and externalists in the philosophy of mind are concerned with the fundamental nature of our thoughts, I mean that they are concerned with the fundamental nature of an important class of our thoughts called propositional attitudes. Propositional attitudes can be construed as composed of an attitude, such as belief, fear, or doubt, and a propositional content, such as that which is believed, feared, or doubted. Specifying an individual's propositional attitudes is essential for action explanation and prediction: if Olivia fears that her tax bill will rise, she may start to save money; if she doubts that it will rise, she may spend more freely. Specifying other people's propositional attitudes and our own is what we do in folk psychology. There certainly are states and thoughts that are not propositional attitudes, but they are not of interest to us, so I will speak of internalism and externalism about mental states, as is more customary.[5]

What mental states an individual is in is causally related to the environment in certain obvious ways: when you cycle through the city, your beliefs about the location and velocity of other objects in your environment constantly adapt. While everyone will agree to this, one may also think that the environment is fundamentally inessential to what mental states you are in in the following way: while the environment may cause your mental states in the actual case, the same mental states could have been caused by something or someone else, a neuroscientist, for example, who stimulates your brain in the right way. A range of views about mental states make the environment inessential to being in a certain mental state: type-identity theories, for example, that identify mental states with brain states, or functionalism that identifies mental states with the functional role they play in an individual's

[5] In this section, I will draw on Jessica Brown's particularly clear way of introducing the fundamental difference between the role that state internalists and state externalists attribute to the environment in determining an individual's mental states (see Brown 2004. Chap.1).

mental system (see, e.g., Lewis 1972; Shoemaker 1982).[6] According to either view, a mind that existed alone in the universe could be in the same mental state as you as long as they were in the same brain state or in the same functional state as you are.

The conception of the relation of mental states and the environment just described is an internalist one. Externalists, in contrast, believe that a mind that exists alone in the universe and a mind that does not exist alone in the universe may be in different mental states even if they are in exactly the same brain states or functional states. This is because, on the externalist view, the environment does not just cause you to have certain mental states, it may partly determine what they are. We can take the way of speaking literally to illustrate the basic externalist intuition. Olga's whole life has been a sham: in her universe, there is an all-powerful demon who is creating the illusion of a physical world external to her but in fact Olga's body is floating alone in a bubble of oxygen in space. The demon is feeding her brain sensory signals to make it look to her as if there is an Earth and people and everything else. In particular, let us say, the demon is feeding her information about what is seems to Olga just like the city of Berlin in our universe: they will make it look to Olga as if she visited "Berlin", that she has seen movies about "Berlin", received postcards from "Berlin". Now if that is the way things are for Olga, then what mental states can she have? Could she think of Berlin in winter that it is charmingly depressing, just as you and I can, for example? There is no Berlin in her universe to have beliefs about. If Olga's belief was about Berlin, it would be true just in case Berlin — that city — is charmingly depressing in winter. But whether Olga's belief is true cannot depend on Berlin as Berlin has never existed in her universe. And how could Olga have acquired the concept of Berlin? Again, there is no Berlin in her world. So, state externalists conclude, Olga does not have a belief, or any mental states, about Berlin, the city.

If Olga's beliefs are only Berlin-beliefs if Berlin exists or has existed in her universe, then Olga's Berlin-beliefs depend on the existence of something

[6] This is not strictly speaking correct: so-called long-arm functionalism individuates functional roles externally. I will come back to it in detail in §2.3.2.2.

— the city Berlin — external to Olga. In the philosophy of mind, the internal/external distinction is understood *metaphysically*: the internal is some metaphysical property of the individual, such as their physical or phenomenal properties. External properties are all those properties that are not their internal properties. If the existence of Berlin is a necessary condition for Olga to be able to have Berlin-related mental states, her mental states do not just supervene on her physical or phenomenal states. Internalism is traditionally formulated as the doctrine that an individual's mental states supervene on the individual's physical or phenomenal states (see, e.g., Burge 1988; Chalmers 1996; Jackson and Pettit 1996; Boghossian 1997; Davies 1998; McLaughlin and Tye 1998; Farkas 2008):

> *State internalism*: An individual's mental states supervene on the individual's physical or phenomenal states.[7]

State externalism is defined as the denial of state internalism:

> *State externalism*: An individual's mental states do not just supervene on the individual's physical or phenomenal states.

I said that propositional attitudes are composed of an attitude and a propositional content. The story about Olga illustrates the basic externalist intuition about propositional contents but we can tell a similar story about attitudes. Let us change the story slightly and suppose for now that Olga has very recently been moved to the demon world. Most concepts Olga thinks with, we use too: when she thinks "full moon", she thinks of a full moon. Like this, we will be able to ignore issues relating to content externalism and focus on attitudes.

In our universe, you sometimes see that it is full moon. Can Olga see that it is full moon? There is no moon in Olga's universe, or anything else she

[7] Whether state internalists will consider purely internal states to supervene on the physical or the phenomenal states of an individual will depend on their preferred metaphysics of mental states.

could see, nor is there any light. It may seem to her that she sees that it is full moon because the demon makes it seem to her that she is but she does not actually see it. If the existence of things outside of Olga is necessary for Olga to think certain contents, it would also seem to be necessary for certain attitudes Olga may take towards these contents.

An externalist about mental states takes an externalist position on both, attitudes and thought contents. I will mostly simply speak of an individual's "*external mental states*" because if either the content or the attitude is externally-individuated, the mental state does not just supervene on the individual's physical or phenomenal states. Conversely, an internalist about mental states takes an internalist position on both, attitudes and thought contents. I will speak of an individual's "*purely internal mental states*" when I mean those mental states that combine an internally-individuated attitude with an internally-individuated content.

Many find the state externalist take on the mind very convincing. It seems plausible, for example, that our conceptual abilities are subject to certain causal constraints, such as that it is a condition of possibility for thinking about certain things that either we or other members of our linguistic community have interacted with that entity at some point. It would seem, for example, that we could imagine a physical and phenomenal duplicate of Olga, who lives in our world and has visited Berlin, seen movies about it, received postcards from it. Olga's duplicate would seem to have all sorts of Berlin-related mental states. As we will see in §1.2, the most influential arguments that state externalists have presented for their views involve cases where we compare two duplicates in different environments that, according to state externalists, are in distinct mental states in virtue of these differences in their environments.

In fact, even many state internalists have found the state externalist arguments convincing, so convincing that they have come to distinguish between a fundamental kind of mental state that supervenes on an individual's internal states, and a derivative kind of mental state that does not exclusively supervene on an individual's physical or phenomenal states. State internalists have labelled the distinction in different ways: they speak of "narrow" or

"epistemic" mental states in contrast to "broad" or "wide" or "subjunctive" mental states (see, e.g., Fodor 1987; Chalmers 2002; Mendola 2008). Many state internalists defend a view according to which an individual's broad states are derivative from an individual's narrow states that supervene on an individual's physical and phenomenal states. We may call those state internalist views "*moderate*" who accept the existence of externally-individuated states. Some state internalists deny the existence of externally-individuated mental states. Such "*hardcore state internalists*" believe that all mental states are narrow states. Unless explicitly stated otherwise, I will be concerned with moderate state internalism.

State externalism also comes in a moderate and in a hardcore version. "*Moderate state externalists*" deny the priority of narrow states (see, e.g., McDowell 1986; Block and Stalnaker 1999) whereas "*hardcore state externalists*" deny the existence of narrow states (Yli-Vakkuri and Hawthorne 2018 are an example of a hardcore view for thought contents).

1.1.2 Internalism/externalism about justification

I said that the internalism/externalism debate in epistemology is concerned with the nature of epistemic justification. More precisely, I said that epistemic internalists believe that whether an individual's mental state that *p* at a time is justified supervenes on what is internal to that individual at that time; epistemic externalists deny that. The internalism/externalism debate in epistemology could not primarily be concerned with the notion of knowledge because knowledge is an externalist notion: whatever sense is given to the "internal" in epistemology, knowledge is not a property that is internal to an individual in virtue of the truth-condition on knowledge. In order for an individual's beliefs to be true, the world has to be how the individual represents it to be.

Still in general, any debate on the nature of epistemic justification is concerned with determining the conditions of appropriate positive evaluation of beliefs with regard to reaching the goal of knowledge. We care about

having justified beliefs because if a belief is justified it is more likely to be known.[8] We may also say that "[un]justified" is generally used in order to evaluate how well an individual's belief is doing with respect to responding to *epistemic* reasons, as opposed to other kinds of reasons, such as prudential reasons. Believing that my teenage children will not drink and bike may be a *prudentially* justified belief because I will sleep better but it is unlikely that it is an *epistemically* justified belief because the fact that I will sleep better does not make it more likely that I know that my teenage children will not drink and bike. More natural synonyms for "[un]justified" are epistemic evaluations like "[ir]rational" or "[un]reasonable", but I will stick to the mainstream epistemological focus on "[un]justified".

Generally, when a person knows some proposition or other, she does so on the basis of something such as adequate evidence or good reasons. The same is true of justified beliefs that may fall short of knowledge. If a belief is justified, it usually is justified in virtue of some further condition(s) obtaining, such as that the individual possesses good epistemic reasons or adequate evidence for the belief or that the belief was produced in a reliable manner.[9] Take Owen who believes that Democrats eat children. Owen's belief will not be supported by the evidence Owen has. Some sources support Owen's belief but they are not reputable. Other people also believe that Democrats eat children but they draw on the same sources. The belief is dismissed by all reputable sources and is extremely unlikely given the prior likelihood of cannibalism in any given group of society. Owen will not be justified in his belief according to any justification theory. I will understand evidence as whatever it is that gives reasons for belief: if there is a difference between reasons for belief and evidence, it does not matter for my purposes.[10] For the sake of presentation, I will speak of evidence in terms of propositions but

---

[8] Truth is often identified as the epistemic goal of justified belief (see, e.g., Silins 2020). But we may have true beliefs by accident and we do not care about those. We care about having stably true beliefs, i.e. knowledge (see, e.g., Sosa 1999; Williamson 2000; Pritchard 2005).

[9] There may be exceptions to this: maybe the rational insight into some simple conceptual or logical truths is so immediate that it is not based on evidence.

[10] There is disagreement about whether things other than evidence can provide epistemic reasons (see, e.g., Foley 1993; Schroeder 2015; Sylvan 2016).

nothing much hinges on that choice.[11] It may be that other things such as states of affairs or mental states (and not just the propositions embedded in mental states) could also constitute evidence. Differences in the metaphysics of evidence will not matter for the discussion of the consistency of hybrid epistemic internalism.

One further clarification before I turn to the internal/external distinction about justification. Nearly everyone agrees that it is not sufficient for a belief's justification that one *has* good reasons or adequate evidence for the belief. In addition, the belief must stand in the right relation to those reasons, or as it is often put, it must be *based on* those reasons if it is to count as justified. This latter requirement is called the "basing requirement" on epistemic justification (for discussion see Korcz 1997). When both conditions are satisfied a belief is *doxastically* justified. Doxastic justification is therefore a property of beliefs. Propositional justification, by contrast, is a property that is had by a proposition relative to a person. A proposition can have such justification for a person even if the person does not believe it or believes it but not for the right reasons. A proposition *p* is *propositionally* justified for a person S just in case S has evidence for *p* such that if S were to believe that *p* and base their belief that *p* on that evidence, their belief that *p* would be doxastically justified. I will be concerned with propositional justification unless stated otherwise.

Now let us come back to Olga who is floating around in her bubble of oxygen in the demon-world. Again let us assume that Olga has only very recently been moved to the demon world in order to focus on issues of justification. We said that everything seems to Olga as it does to us because the demon sees to it. It is clear that Olga falsely believes many things about the world around her. When Olga believes that it is full moon, it will only seem to her that it is full moon because there is no actual full moon. So, Olga does not know that it is full moon. But does Olga have justification to believe that it is full moon? Everything that Olga knows or justifiably believes in her situation strongly suggests that it is full moon: Olga knows or justifiably

[11] See Williamson (2000: Chap.9) for an argument that only propositions can be evidence. For other views on the metaphysics of evidence see, e.g., Pryor (2000) and Audi (2003).

believes that it seems to her that she sees the full moon, that her moon calendar says it is etc. Of course, it is not actually full moon and she does not actually consult her moon calendar, but Olga has no way of knowing that. It would seem to be the only reasonable thing for Olga to believe that it is full moon. It would not be more reasonable, for example, if Olga believed that it was not full moon, even though that would be true, nor would it be more reasonable if she suspended belief on the question whether it was full moon. In her situation, the rational thing for Olga to believe is that it is full moon; so, she has justification to believe that it is full moon.

The conception of justification just described is an internalist one. Even though none of the things that Olga seems to experience are actually the case, she has justification to believe many falsehoods because these beliefs are sufficiently supported by the evidence she has in her situation, namely how things appear to her, and Olga knows or justifiably believes that things appear to her a certain way.[12]

According to *access internalists*, Olga's belief is justified because it is sufficiently supported by facts to which Olga has some relevant access, such as the fact that things appear to her a certain way. BonJour describes access internalism as the "idea that the justifying reason for a basic belief, or indeed for any belief, must somehow be cognitively available to the believer himself, within his cognitive grasp or ken" (BonJour 2002: 24). More precisely, we can define:

> *Access internalism*: Whether or not S has justification to believe that *p* supervenes on facts to which S has some relevant sort of access (see, e.g., Chisholm 1977; Bonjour 1985; Moser 1989; Fumerton 1995; Lehrer 1990; Steup 1999; Audi 2001; Pryor 2001; more recently, Huemer 2006 and Smithies 2019).

[12] To some epistemic internalists, it may be too strong to restrict evidence to only known or even to only true propositions because they believe that justifiably held false beliefs may justify other beliefs (see BonJour 1985; Audi 2001). I will generally speak of an individual's evidence as the propositions they know but everything I will say could be put in terms of justified propositions as well.

Depending on how the relevant epistemic access is specified, the individual may have the relevant access to mental states as well as to non-mental states of affairs. Direct realist theories of perception, for example, hold that external material objects are capable of being directly apprehended in a way that allows their presence to justify the corresponding perceptual beliefs. There is no restriction that justification would only supervene on mental states in the formulation of access internalism itself.

However, other epistemic internalists believe that there should be such a restriction. So-called mentalists have challenged the access internalist reading of Olga's case, although they agree with access internalists that Olga is justified in believing that it is full moon. But mentalists believe that this is not in virtue of what other facts Olga knows but in virtue of the mental states she is in. Mentalists notice that justification would seem to supervene on what is going on "inside" Olga: Olga would seem to have justification for her beliefs even though she is, except for the invisible demon, alone in the universe. Mentalists appeal to an understanding of the internal in an only slightly more metaphorical way than state internalists: according to them, justification supervenes on all those mental states that are "internal" to an individual's mind in some metaphysical sense. For example, conjoined with a materialist view of the mind, justification, according to mentalists, will supervene on some of an individual's brain states. If what matters for justification is that a state is a mental state, i.e. "inside" the individual's mental system, it is negligible whether it can be known by the individual or not. Forgotten or repressed mental states will count toward the justification of a belief, according to mentalists. Mentalists believe that epistemic justification supervenes on all of an individual's mental states at some time whether they are accessible to them or not:

> *Mentalism*: An individual's belief that *p* is justified at some time t if and only if *p* is sufficiently supported by the totality of the individual's mental states at t (see, e.g., Conee and Feldman 2001, 2004; Wedgwood 2002, 2017; Schoenfield 2015).

The formulation of the epistemic externalist position will vary according to the difference between access internalism and mentalism. Epistemic externalists may either deny that individuals necessarily have access to the facts that determine justification, or that justification is necessarily linked to an individual's mental states. They will further allow for external factors, such as causal or counterfactual relations between the individual and the proposition they believe, to partly determine the justification of a belief. Let us define, for reference, in its most general terms:

> *Epistemic externalism*: Whether or not S has justification to believe that *p* at t is partly determined by factors external to the individual at t (see, e.g., Plantinga 1993; Sosa 1999; Goldman 1999; Williamson 2000; Bergmann 2006).

Finally, a terminological clarification. There is one further internalist/ externalist debate within epistemology distinct from the epistemic internalism/ externalism debate: evidential internalism/externalism. Evidential internalism is the view that an individual's evidence supervenes on an individual's purely internal mental states; evidential externalism denies evidential internalism and allows that external factors may partly determine an individual's evidence.[13]

The debates on epistemic internalism/externalism and evidential internalism/externalism are, of course, related. Many epistemologists are evidentialists about justification, i.e. they believe that whether an individual's belief is justified supervenes on the individual's total evidence. In fact, it is often their views on the nature of evidence that turns them into either epistemic internalists or epistemic externalists. For example, many epistemic internalists are epistemic internalists because they are evidentialists on justification and also accept evidential internalism; and many epistemic externalists are epistemic externalists because they are evidentialists on justification and also accept evidential externalism. If hybrid views are possible, epistemic internalism cannot be committed to evidential internalism.

---

[13] See Silins (2005) for framing the distinction in this way.

Hybrid epistemic internalism, the view I will be interested in, usually combines a specific epistemic internalist requirement on justification, like an access requirement, with evidential externalism because it allows for external factors to partly determine an individual's evidence, for example in the form of the individual's externally-individuated mental states.

In sum, justification and evidence are not the same. Some speak of "justification internalism/externalism" instead of the potentially underdetermined "epistemic internalism/externalism" to highlight the difference (Chase 2001; Morvarid 2019). However, I will stick to the more wide-spread terminology of "epistemic internalism/externalism" when I mean internalists and externalists about epistemic justification. Let us have a closer look at those hybrid views next.

1.1.3 Hybrid views: why hybrid epistemic internalism

Hybrid theories are theories that incorporate an internalist or externalist metaphysics of mind and an internalist or externalist epistemology, and cross-combine them. There are two kinds of hybrid theories: hybrid epistemic internalism and hybrid epistemic externalism (**§1.1.3.1**). I will argue that we have theoretical incentives to inquire into the consistency of hybrid epistemic internalism in its two variants, hybrid access internalism and hybrid mentalism (**§1.1.3.2**). I will argue that we do not have a similar incentive to inquire into hybrid epistemic externalist theories (**§1.1.3.3**). Hence, the project of this thesis: assess the consistency of hybrid epistemic internalism.

1.1.3.1 Hybrid epistemic internalism and hybrid epistemic externalism

State internalists and state externalists disagree about whether certain mental properties supervene on an individual's physical or phenomenal make-up. Access internalists and epistemic externalists opposing access internalism disagree about the connection between the justification of one's beliefs and

other accessible states. It would seem that not every state that is internal in the philosophy of mind sense is internal in the epistemological sense: sub-personal states of one's visual processing system, for example, may be internal in the philosophy of mind sense but are arguably not among an individual's epistemically accessible states. Nor is it obvious that every state that is internal in the epistemological sense is internal in the philosophy of mind sense: as was said, we may have the relevant kind of access to non-mental states of affairs. Some have concluded from the fact that state internalists and state externalists, on the one hand, and access internalists and epistemic externalists, on the other hand, are concerned with different conceptions of the internal/external distinction that "it is certainly possible to combine externalism in the philosophy of mind with internalism in epistemology and vice versa" (Pryor 2001: 103).

What about mentalism? We have seen that mentalism explicitly takes inspiration from the way in which the internal/external distinction is understood in the debate about mind. Yet, it would seem as though a mentalist view is also *prima facie* consistent with a state externalist position. This is because mentalists and state externalists are concerned with different constitutional levels of individuals. Mentalism requires that two individuals with the same mental states are justified to the same extent in believing the same propositions. State externalists deny that individuals with the same physical or phenomenal states necessarily are psychologically the same. But the formulation of mentalism does not explicitly say anything about individuals who have the same physical and phenomenal states, only about individuals with the same mental states. So, assuming state externalism, mentalism will not entail that some individuals with the same total physical or phenomenal states have justification to believe the same propositions to the same extent. But that is consistent with the above definition of mentalism. Since the supervenience bases for mentalists and state externalists are located on different constitutional levels, their claims do not contradict each other even though they rely on the same way of understanding the internal/external distinction.

One cross-combination is "*hybrid epistemic internalism*" that was introduced as a view that combines state externalism with epistemic internalism. More precisely, hybrid epistemic internalism formulates specifically epistemically internalist requirements on epistemic justification and allows justification to be partly determined by an individual's external mental states. We can define:

> *Hybrid epistemic internalism*: a form of epistemic internalism according to which, whether an individual has justification to believe that p at t may be partly determined by the individual's external mental states at t.

Importantly, an epistemic internalist theory is not hybrid according to my proposed definition if it merely accepts the existence of externally-individuated mental states but does not allow for them to partly determine whether an individual has justification to believe some *p*. James Chase (2001) proposes a theory along those lines. Chase argues that epistemic internalism is consistent with state externalism in virtue of the fact that it is consistent with the fact that some states are broad that epistemic internalism considers justification to supervene on narrow states and excludes broad states from standing in justifying relations to other states. This is not a hybrid view. Ralph Wedgwood (2002, 2017) and Miriam Schoenfield (2015) propose a somewhat similar version of mentalism about doxastic justification. While they accept externally-individuated states, they argue that factive states could not doxastically justify other states, only their non-factive component could. To set views apart that accept state externalism but exclude external states from entering justifying relationships or of being part of an individual's evidence because they accept evidential internalism, I call these views "*fake hybrid*".

Fake hybrid views are better seen as special versions of the view I call "*pure internalism*". Pure internalist views combine state internalism and epistemic internalism. We may define:

> *Pure internalism*: a form of epistemic internalism according to which, whether an individual's belief that *p* is justified at t is determined by the individual's purely internal mental states at t.

The other cross-combination is "*hybrid epistemic externalism*". Hybrid epistemic externalism formulates specifically externalist requirements on epistemic justification but excludes an individual's external mental states from determining whether that individual's beliefs are justified. We can define:

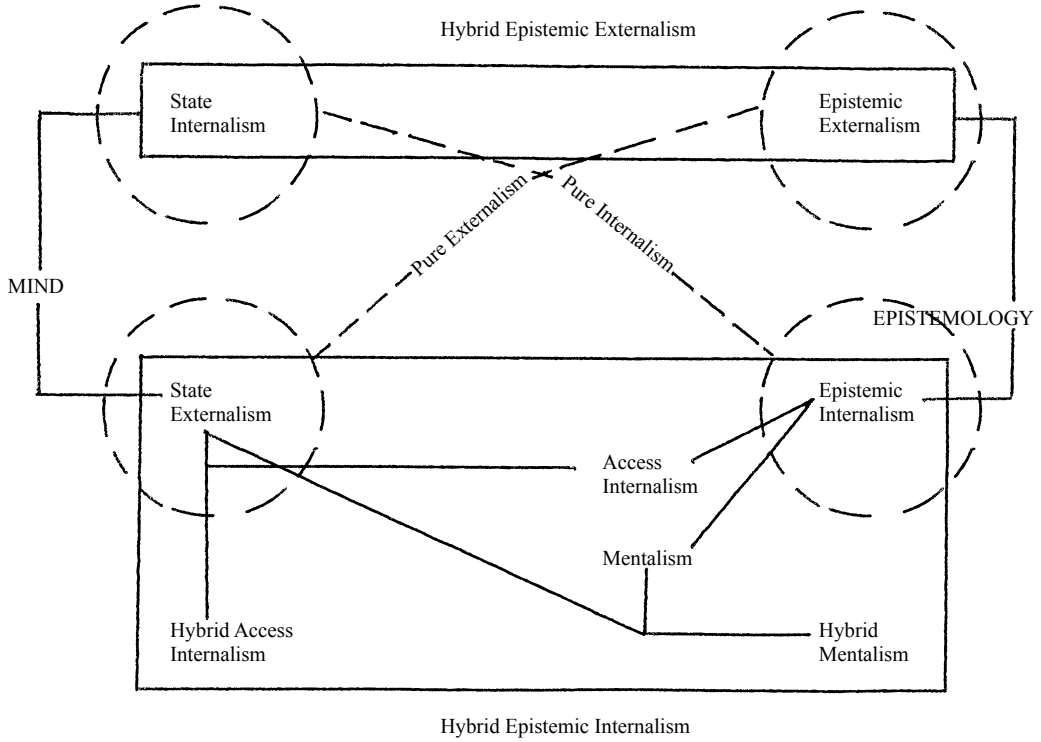
> *Hybrid epistemic externalism*: a form of epistemic externalism according to which whether an individual's belief that *p* is justified at t is *not* determined by the individual's external mental states at t.

This formulation of hybrid epistemic externalism distinguishes it from non-hybrid versions of epistemic externalism without making hybrid epistemic externalism inconsistent by definition. Most (except for certain hardcore) versions of epistemic externalism allow for purely internal mental states to partly determine whether an individual has justification to believe some *p*. They just also allow for external factors to determine whether an individual has justification to believe that p. However, hybrid epistemic externalism could not be the view that *nothing but* an individual's purely internal states determine justificatory status because then, evidently, any external factors would be excluded from determining whether an individual has justification to believe that *p* which contradicts the definition of epistemic externalism.

I will show that while we have theoretical incentives to look into hybrid epistemic internalism, there are no incentives for looking into hybrid epistemic externalism. Among externalist theories of justification, hybrid epistemic externalism contrasts with "*pure externalism*":

> *Pure externalism*: a form of epistemic externalism according to which, whether an individual has justification to believe that p at t is partly determined by the individual's external mental states at t.

Before looking closer at hybrid epistemic internalist views, it may be easier to get an overview of the positions introduced with the help of the following figure:



1.1.3.2 Why hybrid epistemic internalism

As there are two variants of epistemic internalism, access internalism and mentalism, there are two variants of hybrid epistemic internalism. Let us call them "hybrid access internalism" and "hybrid mentalism", respectively. Hybrid access internalist theories hold that mental states do not exclusively supervene on the individual's physical or phenomenal states and that whether the individual has justification for a proposition is a matter of facts to which the individual has the relevant kind of access. We can define:

> *Hybrid access internalism*: a form of access internalism according to which, whether an individual has justification to believe that p at t may be partly determined by the individual's external mental states at t.

Hybrid mentalist theories hold that an individual's mental states do not exclusively supervene on the individual's physical or phenomenal states and that whether the individual has justification for a proposition is a matter of the entirety of the individual's mental states. We can define:

> *Hybrid mentalism*: a form of mentalism according to which, whether an individual has justification to believe that p at t may be partly determined by the individual's external mental states at t.

I will argue that there are theoretical incentives to inquire into both hybrid access internalism and hybrid mentalism.

Common to hybrid access internalism and hybrid mentalism is that they aim to do justice to the fact that state externalism has become an increasingly popular position about the metaphysics of mental states and properties since the first immensely influential arguments for state externalism were presented in the seventies and eighties. I am thinking of the main arguments for externalism about thought content defended by Putnam (1975), Burge (1979), Evans (1982) and McDowell (1984, 1986) that will be introduced shortly (§1.2). Ever since, externalism about thought content has

been extended to externalism about attitudes (see, mainly, McDowell *ibid.,* and Williamson 2000) and even phenomenal properties (see, e.g., Dretske 1995; Tye 1995b; Byrne and Tye 2006). Externalism about mental states and phenomenal properties is supported by externalist theories of self-knowledge: according to these, not only are mental properties partly individuated by the environment an individual is placed in, individuals also have to rely on what they know about their environment to find out what mental states they are in (see Byrne 2018).

At the same time, state internalists have struggled to make internalist accounts of thought contents and attitudes work. We have seen that most state internalists concede to state externalists that some states are partly individuated by the individual's environment. Some content externalists have questioned that any notion of purely internal thought content is viable (see, e.g., McDowell 1986; Adams et al. 1990; Block and Stalnaker 1999; Yli-Vakkuri and Hawthorne 2018). I do not mean to deny that state internalists have developed increasingly powerful opposing theories (see, especially, Chalmers 2002, 2006). Also, state internalism is by no means a negligible position in the philosophy of mind as much in the form of two-dimensional frameworks as well as so-called phenomenal intentionalist views (see, e.g., Lewis 1981; Jackson 1998; Segal 2000; Chalmers 2002; Horgan and Tienson 2002; Farkas 2008). Yet, it is fair to say that inconsistency with state externalism would be a severe cost for any justification theory and may be considered by some to be a deal-breaker against a specific justification theory.

Let us turn to the specific promises of hybrid access internalism and hybrid mentalism, respectively.

Hybrid access internalism combines state externalism and access internalism. Access internalism is attractive because many find it plausible that some epistemic evaluations, such as justification, track what individuals should believe from their epistemic perspective, so to speak. An individual's epistemic perspective would seem to be related to facts to which they have some sort of epistemic access. It seems counterintuitive, for example, that someone could have justification to believe some *p* while none of the things they know or justifiably believe speaks in favour of believing that *p*.

Further, most state externalists believe that individuals have access to the relevant external mental states, i.e. that they are in a position to know that they are in a particular external mental state, because external mental states share certain properties with any mental state (see, e.g., Burge 1988; Falvey and Owens 1994; Laughlin and Tye 1998; Brown 2004). Some have further explicitly defended the view that the kind of access that individuals have to some of their external mental states is of the kind relevant to access internalism: those are either theorists that explicitly defend hybrid access internalist views in some version or other (see McDowell 1986; Gibbons 2006; Pritchard 2012; Das and Salow 2016)[14], or theorists who defend the consistency of state externalism with access internalism in some form (see Audi 2001; Pryor 2001; Brueckner 2002; Madison 2009; Vahid 2003a).

The incentive to inquire into the consistency of state externalism and access internalism derives therefore from the independent attractiveness of both theories, and from the fact that it would seem that individuals have access to their external mental states like to other mental phenomena.

Let us come to hybrid mentalism. The first defence of mentalism as a justification theory by Conee and Feldman (2001) is interestingly intertwined with the internalism/externalism debate about mind. The very first point that Conee and Feldman advance in support of mentalism over access internalism is that it closely mirrors the metaphysical conception of the internal of the debate about mind. They write

> "[T]he root idea is the same. The mind internalist is trying to exclude such plainly external factors as the environmental causal origins and the social milieu of the person's attitudes. Likewise, the epistemic internalist is principally opposed to the existence of any justification-determining role for plainly external factors such as the general accuracy of the mechanism that produces a given belief or the belief's environmental origin. Mentalism bears this out" (Conee and Feldman 2001: 3).

[14] To be fair, not all of the mentioned theorists would self-identify as hybrid access internalists: Das and Salow would consider their view to be an externalist justification theory, as would Gibbons. For discussion see §1.3.1.3-4.

This motivation for mentalism would no longer seem to be available to hybrid mentalists. Hybrid mentalism is the claim that justification supervenes on the entirety of an individual's mental states, amongst others, their external mental states. In that case, justification would seem partly settled by the "belief's environmental origin" or the person's "social milieu".[15]

However, as was later noticed by Conee (2007), mentalism would also seem to be uniquely positioned to be combined with state externalism. Mentalism specifies the supervenience base for justification as the entirety of an individual's mental states *tout court*, without also specifying individuation conditions for mental states. So any mental state, whether an internal or an external state, can in principle be part of the supervenience base for justification according to mentalism. Conee considers hybrid mentalism to be "externally enhanced" by state externalism. He says:

> "So according to mentalism, what is epistemically 'inside of us' is determined in whatever way mental content is determined. Since content externalism expands the factors that fix the mental, content externalism expands the supervenience base for justification according to mentalism. That makes life easier for the epistemic internalist" (2007: 51).

So unlike for hybrid access internalism, the motivation to look into hybrid mentalism mainly derives from the consistency of the compound view itself more than from the independent attractiveness of both theories.

1.1.3.3 Why not hybrid epistemic externalism

There are two main reasons that deflate interest in hybrid epistemic externalism, corresponding to the broader motivations of each of the theories combined. On the one hand, state internalists tend to share a perspective on the individual's mind with epistemic internalists. On the other hand, epistemic externalists have reasons to consider justification to be partly determined by an individual's external states while they lack reasons to consider justification

[15] The question whether justification is actually settled by the "belief's environmental origin" once state externalism is accepted is in fact a more complex one that will be discussed in detail in §2.3.2.1.

to be determined just by an individual's purely internal states. I will comment on each in turn.

State internalists and epistemic internalists alike tend to emphasise individuals' epistemic perspective and tend to link their epistemic perspective to a (supposed) specific epistemic profile of purely internal mental states.

State internalists' motivation for defending the existence, and more importantly, the priority of purely internal mental states is often epistemic. Although accounts of narrow content differ significantly, they are all designed to mirror an individual's epistemic perspective and their patterns of reasoning, in particular how they represent the world to be and which inferences they will be disposed to draw given what they believe to be the case (see, e.g., Loar 1988; Segal 2000; Chalmers 2002). It therefore seems natural for state internalists to endorse a conception of justification that equally emphasises an individual's epistemic perspective (see Chalmers 2002).

Further, state internalists and many epistemic internalists like to link an individual's epistemic perspective to their purely internal mental states in virtue of a (supposed) specific epistemic profile of purely internal mental states. Purely internal mental states are claimed to be epistemically accessible in a peculiar and privileged way that state internalists and many epistemic internalists have characterised, amongst others, as "a priori", "direct", "non-observational" and "incorrigible" (see, e.g., Descartes [1641]/1998; Chisholm 1977; Audi 1998; Bernecker and Dretske 2000). If an individual's epistemic perspective is made up of the propositions that the individual knows in a particularly direct and safe way, and the propositions embedded in purely internal states are always knowable in a particularly direct way, then it seems plausible that an individual's epistemic perspective consists in their purely internal states. Both lines of reasoning lead to what I have called pure internalism: epistemic internalist views that consider justification to supervene on an individual's purely internal states.

It should be stressed though that the pressure for epistemic internalists to adopt state internalism depends on the plausibility of a link between demanding notions of epistemic access, as those just mentioned, and justification. In §1.3.1, I will argue that the cases that epistemic internalists

draw on to motivate epistemic internalism do not support more than a permissive notion of access.

On the other hand, epistemic externalists have reasons to consider justification to be partly determined by an individual's external mental states while they lack reason to consider justification to be determined just by an individual's purely internal mental states in their theory of justification.

One of the main motivations for epistemic externalism is anti-scepticism: most epistemic externalists believe that Olga's physical and phenomenal duplicate in our world has *more* justification to believe empirical propositions than Olga. Assuming state externalism, Olga in the demon world and Olga's duplicate in our world are partly in different mental states. If those external mental states partly determine their respective justificatory status, epistemic externalists have a very natural explanation to give for why Olga and her duplicate are not justified to the same extent in believing the same propositions: they are in different mental states. Conversely, it is because Olga and Olga's duplicate are, by hypothesis, in the same purely internal mental states that epistemic internalists tend to emphasise purely internal mental states in their theory of justification.

An epistemic externalist view according to which an individual's external mental states partly determine an individual's justification would furthermore seem protected against certain complaints that epistemic internalists like to level against any form of epistemic externalism in a way that hybrid epistemic externalism would not seem to be. A typical epistemic internalist complaint against any sort of epistemic externalist view is that it is compatible with epistemic externalism that an individual may be justified in believing some *p* even though they are incapable of knowing that their belief is justified (see BonJour 1985). Compare an epistemic externalist view like Williamson's view according to which an individual's evidence is the totality of propositions they know and knowledge is a mental state, on the one hand, and a hypothetical hybrid epistemic externalist view almost like Williamson's view but that rejects the view that knowledge is a mental state, on the other hand. A view like Williamson's can offer a response to this epistemic internalist complaint. Some epistemic externalists and state externalists have

explicitly defended the idea that it is in virtue of the fact that knowledge is a mental state that individuals have better epistemic access to it than if it was not a mental state (see McHugh 2010; Das and Salow 2016). So on Williamson's view, individuals may be able to know that a proposition is part of their evidence more often than not in virtue of some specific way in which we know that we are in a particular mental state. If an individual's evidence is identified with their knowledge that is not a mental state like on the hypothetical hybrid epistemic externalist view, individuals may fail to know that a proposition is part of their evidence more frequently.

There seems to be little positive incentive to inquire into hybrid epistemic externalism. Pure internalism is considered to be the more attractive package amongst state internalists; and pure externalism amongst epistemic externalists. Hybrid epistemic internalism is the only combination of views that has sparked an interest in hybridity.

Here is the plan for the rest of this first chapter. I will first introduce the component theories in detail. **§1.2** presents the standard arguments for, and different versions of, state externalism. Other than existing research into hybrid epistemic internalism, I will not (myopically) focus on content externalism but will consider whether externalism about the mental in the breadth in which it has been defended for a number of mental aspects is compatible with epistemic internalism. **§1.3** is dedicated to the standard arguments for and different versions of epistemic internalism. I will distinguish between two main motivations for epistemic internalism: the *access motivation* and the *Equal Justification Thesis*. The first, evidently, only motivates access internalism. The Equal Justification Thesis motivates both access internalism and mentalism. In **§1.4**, I turn to the challenges for hybrid epistemic internalism. The two main challenges, corresponding to the two main motivations for epistemic internalism, are the *access problem* and the *equal justification problem*. The remaining chapters of this thesis are dedicated to discussing whether these challenges for hybrid epistemic internalism can be solved, and what follows if not.

1.2 Externalism about Mind

This section has two aims: introducing the major externalist positions about the mental and getting a sense of why externalism about the mind has been thriving for more than forty years (even if not all externalist positions have been thriving to the same extent). The scope of externalism in mind will emerge: almost any feature of the human mind has been "externalised" in some form. I will distinguish between three kinds of externalism: externalism about thought contents (**§1.2.1**); externalism about attitudes (**§1.2.2**) and externalism about phenomenal properties (**§1.2.3**).[16] This section will not provide an assessment of the plausibility of externalism about mind but rather list the considerations in its favour in virtue of which epistemic internalists have felt the need to propose theories of justification that are compatible with externalist theories of the mind.

1.2.1 Externalism about thought contents[17]

Externalism about the mind was first developed as externalism about thought contents, which is still the externalist view about mind dominantly discussed in the literature. Externalism about thought contents comes in three forms, depending on what kind of content it has been defended for: *natural kind externalism* claims that an individual's thoughts are partly individuated by the natural kinds in their environment (**§1.2.2.1**); *social externalism* claims that an individual's thoughts are partly individuated by the linguistic practices of their community (**§1.2.2.2**); finally, Russellian and neo-Fregean *singular externalism* claim that an individual's thought contents are partly individuated by the particular objects in their environment, even if they differ on how this is achieved (**§1.2.2.3**).

---

[16] One may miss so-called vehicle externalism from that list, the view that an individual's thought vehicles can be located outside the physical boundaries of their body. But vehicle externalism is an externalism about the physical substrate of the mind, and not the mind itself.

[17] I owe the taxonomy of content externalist views discussed in this section to Jessica Brown (2004: Chap.1).

1.2.1.1 Natural kind externalism

In "The Meaning of 'Meaning'", Putnam argues for what has been called *natural kind externalism* (Brown 2004). Natural kind externalism is the position that natural kind terms (such as "water" and "gold") mean what they do because we interact causally with the natural kinds that these terms are about. Natural kinds include physical kinds, chemical substances and biological kinds. Natural kinds are individuated by their fundamental properties, as described by correct scientific theories (Putnam 1975; Kripke 1980). It is both necessary and sufficient for an item to be a member of a natural kind that it has the relevant fundamental properties.

Putnam's original aim in "The Meaning of 'Meaning'" (1975) was to dispute certain 'grotesquely mistaken' views of language which, Putnam suggests, depend on two incompatible assumptions about meaning:

(i) The meaning of our terms (for example, natural kind terms) is fixed by the psychological states of those who use them.
(ii) The meaning of such terms determines their extension: a difference in extension is sufficient for a difference in meaning (Putnam 1975: 219).

Putnam presents his famous Twin Earth story as an argument that both assumptions cannot be true of meaning. Instead, Putnam will suggest abandoning the first while retaining a modified form of the second. Here is the well-known story. Putnam asks us to imagine a planet, "Twin Earth", which is just like Earth except that the liquid that runs in rivers and lakes, and quenches thirst and looks and feels in any other respect like water is not H_2O but XYZ. People on Twin Earth coincidentally also call XYZ "water" (we can call it "twin water"). He also supposes that each of us has a type-identical twin on Twin Earth who has the exact same history of purely internal states. You and your twin have exactly the same physical microstructure, your bodies perform the same movements, you have the same patterns of stimulation on your retinas, and so on. Further, we may suppose that you and your twin are in the same phenomenal states as the situations on Earth and on Twin Earth will

seem just the same to you, for the only difference between the two situations is in the fundamental chemical nature of the stuff called “water”.

Now suppose that a person called Oscar on Earth says “seawater is not drinking water” and Twin Oscar on Twin Earth makes the same noises. Do they utter sentences with the same meaning? Putnam argues that they do not since the references of their utterances of “water” are different: H_2O on Earth and XYZ on Twin Earth. Since the extensions of the two utterances of “water” differ, their meaning must differ in virtue of (ii). Putnam insists that the difference in meaning of “water” on Earth and Twin Earth does not depend on the fact that some scientist on each planet could tell the difference between H_2O and XYZ. To illustrate this, he describes the two planets in 1750 before the development of adequate chemistry. In this case, no one could tell the difference between the two substances; but the extension of “water” on each planet differs. Since Oscar has never been exposed to any water lookalike, Putnam believes that Oscar exclusively refers to H_2O by “water”, while Twin Oscar, having never been exposed to any twin water lookalike, refers exclusively to XYZ by “water”. If that is correct, Oscar’s belief expressed by the sentence “seawater is not drinking water” is true if and only if humans cannot drink H_2O with a high concentration of sodium, while the belief Twin Oscar expresses by the same sentence is true if and only if twin humans cannot drink XYZ with a high concentration of twin sodium. So if Oscar visits Twin Earth and says “This is seawater” about the Twin Mediterranean, he will say something false, and vice versa for Twin Oscar. But this is compatible, Putnam argues, with supposing that Oscar and Twin Oscar are physically and phenomenally identical. Since the only difference between the two situations is in the chemical structure of the substance the population on each planet calls “water”, natural kind externalists conclude that the meaning of an individual’s utterances involving natural kind terms is not wholly determined by the

individual's psychological states, contra (i), but is instead individuated partly by the natural kinds in their environment.[18]

The argument only establishes that psychological states do not determine meaning if Oscar and Twin Oscar are in the same psychological states. The states that are not affected by the difference between H_2O and XYZ are narrow states which Putnam permits the existence of (1975: 220).[19] So Putnam's argument really only establishes that narrow states do not determine meaning, but broad states may: the meaning of our terms may be fixed by the broad states of those who use them. So, condition (i) can be kept if modified to a version about broad states. How is the content of broad states determined? Here the reasoning behind (ii) can be extended to broad states. The content of a mental state determines its truth-conditions. A difference in truth-conditions is sufficient for a difference in content. If the truth-conditions of Oscar and Twin Oscar's states differ, so do their contents and so do the states themselves (that are individuated by their contents). If that is correct, Putnam's argument shows that propositional attitudes are broad (and will fix the broad meaning of our utterances).

Putnam's Twin Earth scenario was immensely influential. Not only because of the discussion it sparked about what it shows, but because it presents a method by which to construct a type of thought experiment or case that I will call a "*twin case*". As we will see, twin cases play a crucial role in extending externalism from thoughts involving natural kind terms to almost all mental phenomena, and turning externalism into the dominant view on the nature of the mind that it is today.

Define a case as as a total state of a possible system consisting of a subject *S* in a mental state *M* paired with an environment *E* at a time *t*. We may then define twin cases as follows:

---

[18] Like Putnam, I have ignored the fact that Oscar's body is largely made up of water. This does not significantly affect the debate for Putnam could have used a natural kind that is not found in human bodies. However, Katalin Farkas considers this indifference to reveal that physical definitions of the state internalist/externalist debate are misguided. I agree with Farkas. For more on this see §4.2.1.1.

[19] In the early papers Putnam treats psychological states as narrow – his conclusion is only about linguistic meaning (see, e.g., Putnam 1967; 1975).

Twin cases:

*Take case α and (real or counterfactual) case β: S_α is in M_α in α and S_β is in M_β in β;

*Case α is a *twin case* to case β, if and only if S_α and S_β are physically or phenomenally identical (or both).

We will see that any form of state externalism can be formulated by the help of a twin case, i.e. of a case where we are comparing two duplicates in different environments that, according to state externalists, will be in distinct mental states in virtue of differences in their environments. In other words, the state externalist hypothesis, in its most general form is that S_α in α may be in M_α while S_β in β may be in M_β, and $M_\alpha \neq M_\beta$.

Let us turn to Burge who defends a first significant broadening of state externalism by the help of an equally famous twin case.

1.2.1.2 Social externalism

In "Individualism and the Mental," "Two Thought Experiments Reviewed," "Other Bodies," and elsewhere, Tyler Burge argues for social externalism (Burge, 1979, 1982a, 1982b). Social externalism is the position that our thought contents depend essentially on the norms of our social environments.

Here is Burge's famous arthritis twin case to this effect (Burge 1979). Suppose that Olivia has suffered arthritis for a number of years, and so she has many attitudes about arthritis. Amongst them, she also has the belief that she would express by saying, "My arthritis has spread to my thigh". This attitude indicates that Olivia incompletely understands "arthritis", for, by definition, arthritis is a rheumatoid ailments of the joints, and of the joints only. Despite this, Burge argues, Olivia has the concept ARTHRITIS. According to Burge, by her utterance "My arthritis has spread to my thigh", Olivia expresses her belief that her arthritis has spread to her thigh which is false. Burge supports this interpretation by saying that it would be natural to report her thoughts in this way, despite her incomplete understanding.

Now consider a counterfactual situation in which Olivia is brought up in a different linguistic community in which the expression “arthritis” has a different definition. Whereas in the actual situation, ARTHRITIS is defined to apply to rheumatoid ailments of the joints, in the counterfactual situation, it is defined to apply to rheumatoid ailments of the joints and muscles. In the counterfactual situation, according to Burge, Olivia has no attitudes about arthritis because she does not have the concept ARTHRITIS in that scenario. But she has a different concept, she would express with the word “arthritis”, which we may call THARTHRITIS. In this counterfactual case, Olivia’s belief is true. In support of this interpretation, Burge points out that the experts in the counterfactual community would explain “arthritis” as a rheumatoid ailment of the muscles and joints, and Olivia presumably intends to use the term in the way these experts use it. Despite the difference in how “arthritis” is defined in her linguistic community, Olivia in the counterfactual scenario is stipulated to have precisely the same history of purely internal states than she has in the actual situation. At any time, Olivia is in exactly the same physical and phenomenal states, for the only difference between the two situations is in the way “arthritis” is defined by the linguistic community, of which Olivia is ignorant. Social externalists conclude that Olivia’s states are not wholly determined by Olivia’s physical or phenomenal states but are partly individuated by the uses and definitions of concepts in her linguistic community.

Burge’s thought experiment differs from Putnam’s in two ways. First, it is directly concerned with propositional attitudes. Second, Burge does not restrict his externalism to natural kinds, or even to obviously social kinds, for it can be applied to any term that an individual could misuse.

1.2.1.3 Russellian and neo-Fregean singular externalism

Singular externalists hold that some states are object-dependent in the following sense: some states are such that they are essentially, existentially or both related to their intentional object (see, e.g., Perry 1979; Kripke 1980; Evans 1982; Peacocke 1983; McDowell 1984; Salmon 1989; Campbell 1987;

Kaplan 1989). Common to all forms of singular externalism is that they concern the individuation of singular thoughts where we can define singular thoughts as those thoughts that would be expressed by a sentence containing a demonstrative, a proper name or an indexical, i.e. expressions commonly considered to refer to particulars directly and non-descriptively. Singular thoughts, singular externalists claim, are individuated partly by the particular objects that they are about. Singular externalism comes in two variants: *Russellian* and *neo-Fregean* singular externalism.[20] While Russellian and neo-Fregean singular externalists agree on the object-dependence of singular thoughts, they disagree about how the object-dependence is achieved.

On the standard interpretation of Russellianism we can think about an individual directly by having that individual as an immediate constituent of the thought. For a contemporary version of this basic Russellian idea, take relationalists (also called naïve realist) views about perceptual content. Relationalists believe that perceptual experiences are complex, relational events partly consisting in the presentation to subjects of external objects and their qualities (Martin 1997). Suppose that, in the actual situation, Oliver is looking at a certain apple, A_1, and he thinks that that apple is overripe. In this situation, his thought refers to the particular apple, A_1, he is looking at, and its truth-value turns on the state of that apple, A_1, that is on whether it is overripe. Now consider a counterfactual situation in which everything is the same, but Oliver is looking at a distinct apple, A_2, which looks just like A_1. In the counterfactual situation, Oliver is in exactly the same physical and phenomenal states. However, in virtue of the fact that he is looking at the different apple A_2, his thought refers to A_2, and it is A_2's state on which the truth value of his thought depends. The difference between the two situations consists entirely in the different apples that Oliver is seeing; and so does the difference between the two mental states. Or take a case where Oliver first sees an apple and, in a counterfactual situation, hallucinates an apple. In the twin case, Oliver does not see a different apple, he does not see an apple at all,

[20] Famously, Russell himself held a descriptive theory of proper names, at least insofar as they occur in thought contents or idiolects (it may be that Russell held a distinct theory about names in public languages that is closer to a singular externalist view, see Sainsbury 1993).

so he is not related to one. Consequently, according to Russellian singular externalists, Oliver's perception of an apple and his perfectly matching hallucination of an apple cannot be the same mental state (Martin 2004, 2006).

On Russellian singular externalism, the object-dependence of singular thoughts follows straightforwardly from the object-involvement in singular thoughts. Naturally, Russellian singular externalists conclude that Oliver's states are not wholly individuated by his physical or phenomenal states, but are partly individuated by the objects in his environment.

Famously, Frege had a different conception of propositions, including those embedded in singular thoughts. On the standard interpretation of Frege, individuals could not be constituents of propositions as propositions are composed of senses. Frege famously distinguished between the object, or referent, of a thought, and the way the subject thinks of the object, the sense (Frege 1948/originally 1892). The sense, according to Frege, is the content of a thought. To take a classic example, assume that an early astronomer makes observations of the planet Venus in both the morning and evening but incorrectly takes the morning and evening observations to be of different stars. They coin two terms—"Hesperus" and "Phosphorus"—for what they regard as two distinct stars, one visible in the evening and one in the morning. An expression's sense, according to Frege, is intended to capture its *cognitive value* where the cognitive value of a thought is thought to explain a number of an individual's epistemic behaviours, such as which identity judgments they will find informative and which inferences they will be inclined to draw. Frege proposed that the astronomer thinks of a single object, Venus, in two different ways, or via two different senses.

Thinking back to what I said about the function that the notion of narrow content is generally thought to fulfil, namely mirroring an individual's epistemic perspective in terms of how they represent the world and how they would reason about it, it is unsurprising that Fregean senses traditionally belonged to the state internalist's toolkit for specifying mental states. This is made explicit in the orthodox reading of Frege, which holds that Fregean senses are descriptive and therefore existentially and essentially *in*dependent of the objects they are about. The thoughts the astronomer would express with

"Hesperus", for example, would, according to this view, involve some descriptive component of the form "the star that is visible in the evening". In that case, their thought would be essentially and existentially independent of any specific object because they would think of whatever object uniquely fits the description, if there is one. On this understanding of sense, one cannot combine the idea that an individual's thought is individuated partly in virtue of the particular object the thought is about with the idea that individuals think of objects under a specific sense. Some have therefore argued that content externalism and Fregean senses are incompatible because they assume that whenever a subject thinks of something via a sense, they think of it via some description (see, e.g., Kripke 1980; Salmon 1989).

So-called neo-Fregeans like McDowell and Evans, however, propose to think of Fregean senses differently, namely as themselves being object-dependent. So-called *de re senses* are semantic units that, without involving the particular itself as a constituent, are such that they would not be available to be entertained if the particular did not exist (McDowell 1984: 204).

We can apply neo-Fregean singular externalism to perceptual demonstrative thoughts. Take the perceptual demonstrative thought that that cat is sleepy. On neo-Fregean singular externalism, the demonstrative component of this thought is not exhausted by the particular cat being referred to. Instead, its content is given by the object-dependent sense or way of thinking about that cat. The thought that that cat is sleepy contains, in subject position, a mode of presentation of that cat. But that mode of presentation itself owes its existence and/or identity to the particular cat. That sense could not have presented a different object, nor could it have failed to present any object at all.

Now compare a counterfactual situation in which everything is the same, but you are looking at a lookalike cat, a sister cat let's say, and think that that cat is sleepy. In the counterfactual situation, you are in exactly the same physical and phenomenal states. However, in virtue of the fact that you are not looking at the same cat, the sense of thinking of the first cat is not available when you are looking at the sister cat. So you cannot think the same thought as in the actual case. On neo-Fregean singular externalism, the object-

dependence of singular thoughts follows from the object-dependence of the senses of singular thoughts. It follows that your states are not wholly individuated by your physical or phenomenal states, but are partly individuated by the objects in your environment.

1.2.2 Externalism about attitudes

Consider again naïve realist views that consider perceptual experiences to be complex, relational events consisting in the presentation to individuals of external objects and their qualities (Martin 1997). In virtue of the fact that veridical perceptual experiences present external objects and their qualities to the individual, they could not belong to the same mental kind as their matching hallucination according to naïve realism. In fact, naïve realists typically claim that veridical perceptual states have no "highest common factor" with hallucinations (McDowell 1983: 386). Naïve realism is thus a *disjunctivist* view about perceptual experiences: according to it, the category of perceptual experience, which is supposed to subsume three types of perceptual experiences (veridical experiences, illusions and hallucinations), is a heterogeneous or *disjunctive* category.

What we may want to call "*anti-conjunctivism*" in epistemology is the view that epistemologically successful states, such as seeing that something is the case, are not compound states, built up out of epistemologically failed states. Williamson's knowledge-first view is a prototypical anti-conjunctivist view: knowledge is basic and cannot be constructed out of truly believing something in a justified way (Williamson 2000: Chap. 1-2).

Common to disjunctivism and anti-conjunctivism is that they distinguish between successful states on the one hand and failed states on the other – even though the two kinds of state may be indistinguishable for the one who instantiates them. The successful states include seeing, hearing (or 'could hear', i.e. hearing not in the sense of testimony), feeling (in the sense of touch) etc. that something is the case. The failed states include merely seeming to see or hear or feel etc. that something is the case.

The distinction between successful and failed states is a distinction that pertains to the *attitude* one takes towards some content: the content could be the same between a successful and a failed state. The external factors therefore play a role in partly determining the type of attitude that the individual takes toward some content. For this reason, disjunctivism and anti-conjunctivism are considered to be forms of *attitude externalism* (Williamson, 2000: chap. 2). Attitude externalists believe that some attitudes are *factive*: they entail the truth of the content represented; whereas others are non-factive and compatible with the truth and falsity of the content they are directed at (see, e.g., McDowell 1983; Williamson 2000).

If we accept the existence of factive mental states such as seeing, hearing (not in the sense of testimony) and feeling (not in the sense of touch) that something is the case, it can seem like a small step to generalise from there and claim that knowledge is a mental state, i.e. the state of knowing that something is the case. For Williamson, knowledge naturally takes its place as the most general factive state – the factive mental state that one is in when one is in any specific factive mental state (Williamson, 2000: section 1.4). On Williamson's view, it is not possible that two individuals agree in all mental respects, yet one of them has knowledge, while the other does not. Whether I know that I had oranges for breakfast, that Macron paved the way for right-wing populism in France or that there is a black hole in the middle of our galaxy, my mind is just in a particular state, according to Williamson's view.

1.2.3 Externalism about phenomenal properties

A highly influential view on the phenomenal character of mental states is representationalism. Representationalism is the view that the phenomenal character of an experience is fixed by some or all of the experience's representational features, its representational content (see Dretske 1995; Tye 1995; Lycan 1996). The phenomenal character of my visual experience of a lime, for example, is to be understood entirely in terms of what my experience represents the lime to be. There are views of various strength about what this

"is fixed by" relation amounts to. Some representationalists specify it in terms of a supervenience relation; others in terms of identity. The central representationalist claim about phenomenal character is the following:

> *Phenomenal representationalism*: For every phenomenal property P, there is a propositional attitude A to the proposition *p* such that P is identical to (or supervenes on) the property of having A to *p*.[21]

If content externalism and representationalism about phenomenal character are conjoined, they appear to lead to externalism about phenomenal character: if an experience's phenomenal character is or supervenes on its representational content, and if the representational content is externally-individuated, then the phenomenal character may itself be externally-individuated. Tye, a leading externalist representationalist, writes for example:

> "The ... phenomenal character of a perceptual experience consists in, and is no more than, the complex of qualities the experience represents. Thus, the phenomenal character of the experience of red just is red. In being aware of red, I am aware of what it is like to experience red, since what it is like to experience red is simply red ... The phenomenal character of an experience, then, is out there in the world" (2009: 119).

Assuming naïve realism about perceptual experiences, the case generalises to more complex phenomenal characters. Think of Oscar and Twin Oscar. Oscar may learn how to recognise water, he can tell water from gin, oil, and other liquids. In good circumstances, he can tell just by looking that there is water in front of him. Twin Oscar can do the same things about twin water. A naïve realist, who accepts that there are perceptual recognition capacities for natural kinds, will think that there is an aspect of Oscar's perceptual experience which is different from that of Twin Oscar's: Oscar judges water present by looking and Twin Oscar twin water.

Importantly, naïve realists believe that this is compatible with the fact that water and twin water share all their appearance-properties: Oscar and

---

[21] This is a simplified version of what Speaks proposes (2015: 467).

Twin Oscar may realise that water and twin water, respectively, look a certain way and that something could look just like that even if there was no water or twin water, respectively, present.

So naïve realists will believe that there is more to Oscar's and Twin Oscar's perceptual experience than what their perceptual experiences have in common. But, importantly, while water and twin water are not phenomenally identical, they share some phenomenal properties (phenomenal externalism would be an absurd view if it denied that).

The same remarks extend to whatever conscious phenomenal character hallucinations or illusions have, according to naïve realists. Even a perfectly matching hallucination does not have to share all its phenomenal properties with a certain veridical perception although they will, of course, share certain appearance-properties.

Naïve realism, then, is an example of a view where the content, the attitude and the phenomenal properties of certain perceptual states are dependent on the environment. In other words, even if two individuals are in the same physical states, things will not appear the same to them according to phenomenal externalism. Note that, evidently, phenomenal externalism is not compatible with a characterisation of twin cases in phenomenal terms.

In sum, what this section shows is that what has started as a view on the meaning of natural kind terms has been developed — by the help of twin cases — into a consistent outlook on every aspect of the mind.

1.3 Internalism in Epistemology

This part is dedicated to the central cases and considerations that epistemic internalists have put forward in support of their views. I will distinguish between two main motivations for epistemic internalism: the *access motivation* and the *Equal Justification Thesis*. The former states that there is a link between whether an individual's belief that *p* is justified and certain facts to which the individual has the relevant sort of access. The epistemic access

that is usually considered to be relevant to access internalist views about justification is *special access*. Here I will argue that the cases and considerations behind the access motivation only support a link between justification and a more permissive sort of access that I will call "*ready access*" (**§1.3.1**). The Equal Justification Thesis is the view that individuals in what I will call, following Williamson, "*bad cases*" have justification to believe the same propositions as their internal duplicates in the "*good case*" (2000: Chap.8). As we will see, it is in order to vindicate the Equal Justification Thesis that access internalists consider special access to be the access relevant to justification (**§1.3.2**). This will become crucial in later chapters and in the overall outlook of this thesis. In chapter 2, I will argue that the Equal Justification Thesis in its original form cannot be vindicated by hybrid epistemic internalism, that is once state externalism is assumed. But if the Equal Justification Thesis cannot be kept, hybrid access internalists might as well focus on the more plausible sort of access, namely ready access. This thesis is also a systemic investigation into how internalist positions in both mind and epistemology can be formulated without reference to special access, and into the shape their disagreement with externalists takes in that case. One of the main results of this thesis is that not only would most points of conflict between internalists and externalists arise in a very similar fashion if put in terms of ready access instead of special access, but it would seem that we get a *better* understanding of the internalist/externalist disagreement if it is seen to focus on the significance of what is readily accessible to individuals.

1.3.1 The access motivation

I said that we care about having justified beliefs because if a belief is justified it is more likely to be known. But, access internalists point out, certain case intuitions are at odds with this claim. An essential property of knowledge is factivity: when we know things, we got things right and this non-accidentally. But when we evaluate people's beliefs as justified, it would seem to matter more whether individuals respond to certain facts to which they have some

relevant kind of access rather than whether they achieve knowledge. The first two famous internalist case intuitions that I will consider, Norman, the clairvoyant (**§1.3.1.1**), and the New Evil Demon problem (**§1.3.1.2**), support, according to access internalists, the view that justification is more closely linked to what individual have the relevant sort of access than to whether they achieve knowledge. I will further argue that a certain interpretation of Norman's case as well as further general considerations about how we epistemically criticise each other's beliefs suggest that the relevant notion of access is a permissive form of access which I will call "ready access" (**§1.3.1.3**). Finally, further considerations about the social practice of justifying our beliefs to others have been considered by some access internalists to support that justification would require access not only to the facts that determine one's justificatory status but to the status itself (**§1.3.1.4**).

1.3.1.1 Norman, the clairvoyant

Norman is, under normal conditions, a completely reliable clairvoyant. He possesses no evidence or reasons of any kind for or against the general possibility of a clairvoyant ability or for or against the possibility that he possesses it. One day Norman comes to believe that the President is in New York City, though he has no evidence either for or against this belief. In fact, the belief is true and results from his clairvoyant power under circumstances in which it is completely reliable (Bonjour 1985: 41). But from Norman's perspective this belief seems like a random hunch of which he finds himself oddly convinced. BonJour concludes that despite the fact that Norman's belief is true and formed in a perfectly reliable manner, Norman does not have justification to believe that the President is in New York.

From cases like this, Bonjour derives the conclusion that some kind of access to something that speaks in favour of one's belief is required on the part of the individual in order to be justified, and that reliability is not sufficient for being justified. More precisely, Norman's case suggests that justification minimally requires *some* access to *some* reasons for thinking that one's belief is true. This is because Norman lacks any access to any such reason: from his

point of view, he has no reason to think that the President is in New York, nor is he aware of any possible way he could come to have a true belief on that topic. Our conception of justification does not seem compatible with cases of justified belief where the belief seems arbitrary to the individual. Given the standard access internalist reading of Norman's case, the case suggests that justification requires access *tout court* to the fact that one's belief has something going for it, so to speak; *contra* reliabilist views of justification.

However, there is also a non-standard reading of Norman's case that has been put forward by singular externalists (Dancy 1985; McDowell 1998) and has been dismissed by BonJour (BonJour 1985: Chap.4; BonJour 1992: 35). The non-standard reading, I believe, in fact nicely supports access internalism and gives insight into the kind of access that may be linked to justification.

According to the standard *inferentialist* interpretation of Norman's case, we are asked to picture Norman's epistemic situation as follows: Norman finds himself with some appearances of the president being in NYC that day, and will try to infer from some reasons that these appearances are accurate. Given that he does not have any reasons to believe that the president is in NYC from another trusted source, nor any evidence about possessing a clairvoyance ability, it would seem hopeless for Norman to trust these appearances. But we can also picture Norman's clairvoyant ability to be like a perceptual ability that he did not know he possesses and that allows him to directly perceive certain bits of reality. Just like we would see that a window is open upon entering a room, Norman would 'clairvoyant' that the president is in NYC.

According to this non-standard interpretation, Norman would have *non-inferential* justification for his clairvoyant beliefs, just as many foundationalists about justification accept we have for other perceptual beliefs.[22] Given this way of picturing the case, Norman would have access to a reason supporting his clairvoyant belief that the president is in NYC, namely

[22] Foundationalism about justification is the view that all justified beliefs rest ultimately on a foundation of non-inferentially justified beliefs. Foundationalism in its classical internalist version has been defended by Descartes ([1641]/1998) and Russell (1910); in an externalist version by, for example, Goldman (1979).

the reason that he 'clairvoyants' that the president is in NYC (although maybe not under this description: Norman may have ready access to the fact that he 'sees' it in some way or just knows it). Consistently with access internalism, the intuition that Norman is not justified in believing that the president is in NYC is much harder to uphold under the non-standard interpretation.

Assuming the non-standard reading of Norman's case, the case would furthermore provide insight into the kind of access justification may be linked to. To 'clairvoyant' that the president is in NYC is a presumably a sort of direct and empirical access Norman has to the state of affairs that the president is in NYC. So Norman would seem to have some direct, empirical access to the proposition that is part of his evidence. To be sure, such a direct realist theory of (clairvoyant) perception would be rejected by most contemporary access internalist foundationalists. According to these access internalists, individuals have the relevant sort of access only to propositions about how things seem or appear to them (see Pryor 2000; Huemer 2007). Further, Norman has some sort of direct access to the fact that the proposition that the president is in NYC is part of his evidence: he knows that he 'clairvoyants' that fact because he 'clairvoyants' it. I will argue below (§1.3.1.3) that limiting the access relevant to access internalism in the way in which Pryor and Huemer propose to do, contradicts the notion of access that is intuitively linked to justification.

1.3.1.2 The New Evil Demon (NED) Problem, 1st take

While Norman's case seems to illustrate that reliability is not sufficient for justification, epistemic internalists claim that bad cases provide evidence that reliability is not even necessary for justification. They defend this claim by appealing to what has become known as the New Evil Demon problem (Lehrer and Cohen 1983; Cohen 1984).

We are asked to imagine a demon world, like the one in which Olga lives. The demon-world is a typical bad case: in the bad case, things appear as they ordinarily do but are some other way (see Williamson 2000: Chap.8). If

an individual in a bad case believes an empirical proposition, say that they have hands, it will generally be false and not be known.

Now Cohen asks us to imagine two inhabitants of the demon world, let us call them Barry and Warren. Barry is careful, epistemically speaking: he reasons in accordance with good epistemic standards in response to the evidence he has. Warren, on the other hand, is epistemically reckless: he engages in confused reasoning and wishful thinking, has biases and sometimes simply guesses what to believe. For further reference, we may suppose that Barry has a physical and phenomenal duplicate Gary who lives in the good case. In the good case things appear as they ordinarily do and are that way. If someone in the good case follows good epistemic standards, they generally come to know what they believe. In fact, we can define a good epistemic standard as those standards that, generally, lead to knowledge in the good case. For example, if Gary believes that he has hands because he sees them, he knows that he has hands. I will come back to Gary; for now let us focus on Barry and Warren.

We can imagine that the demon makes it seem to Barry and Warren as if they come across an intricate crime scene and try to figure out what happened. Warren will confabulate some unlikely story unrelated to how things appear to him whereas Barry will not miss the smallest piece of apparent evidence, will consider the complete array of plausible hypotheses and will settle for the one seemingly best supported. Of course, in fact, both are flying around in deep space alone with no means to communicate. Any of their respective beliefs is equally unrelated to how things actually are. But it is clear that Barry does something better than Warren: he is highly responsive to the evidence he has and reasons carefully. Cohen suggests that the difference between Barry and Warren is marked precisely by the concept of justification: Barry is *justified* in his beliefs whereas Warren is not justified (1984: 283).

The New Evil Demon problem is a problem for those who believe that knowledge, and relatedly, truth matter more with respect to whether someone has justification for a certain belief than whether they respond to the evidence that is available to them, however deprived their epistemic situation may be. Barry's case would seem to suggest that as long as someone conducts their

epistemic affairs properly, so to speak, they may have justification for certain beliefs even if they form reliably false beliefs. The lesson access internalists draw from comparing Barry to Warren is that we use the notion of justification to evaluate how well someone responds to certain facts to which they have access in a context rather than whether they getting closer to achieving knowledge.

Note that, similar to the access suggested by the non-inferentialist reading of Norman's case, the kind of access relevant to justification would seem to be some sort of fairly direct access: Barry does, and Warren does not, rationally respond to how the crime scene appears to him, what the apparent probabilities are about how other apparent people behave in the demon-world etc. The fact that Barry has fairly direct access only to how things seem to him whereas Norman has fairly direct access to how things actually are would seem to be negligible with respect to whether they have justification for their beliefs in their respective contexts.

1.3.1.3 Epistemic criticism and ready access to reasons

One way to put the difference between Barry and Warren is that Barry believes as he should believe whereas Warren does not believe as he should believe, epistemically speaking. More generally, access internalists like to stress that we criticise *people* for beliefs that they should not have, and sometimes, even if less often, approve of the fact that they believe as they ought to (Ginet 1975; Steup 1999; Gibbons 2006). The fact that we criticise people for their beliefs suggests that we use "[un]justified" as an *agent-directed* evaluation. Agent-directed evaluations target individuals for something they did or failed to do; it is their answer (or failure to answer) with respect to some standard that is at issue, and not just whether some standard is met (Kiesewetter 2017: 34). We can evaluate a knife, for example, as meeting the standard of a good knife and, clearly, this is not an agent-directed evaluation. We could evaluate a belief that *p* purely with regard to whether it meets the standard of a rational or justified belief, but, access internalists emphasise, we would seem to evaluate whether the individual meets the standard of rationally believing that *p* if they

believed that p given the evidence they have. Some theorists invoke the distinction between "deontological criticism" and "normative criticism" to mark the difference between agent-directed and purely norm-directed evaluation. The issue is that the notion of deontology is associated with all sorts of further notions such as duty and blame that elicit strong intuitions that may not be helpful in the current context.[23] The point that access internalists like to stress and that can be stressed without invoking charged notions such as duty and blame is that "[un]justified" is used to evaluate whether people did something well or not with respect to their beliefs, or, as I said, whether they are conducting their epistemic affairs properly.

Now access internalists suggest that if justification is an agent-directed evaluative notion, some form of access internalism must be true. The link is supplied by an epistemic "ought implies can" principle (Ginet 1975): if an individual should believe something, then they could believe it; if they should not believe something, then they could refrain from believing it. Since beliefs are not subject to voluntary control like actions, the "can" in question is a "can" of rational abilities.[24] Crucially, the "can" of rational abilities that access internalists consider to be at stake has two layers: in any specific context, the individual knows or justifiably believes what their evidence is, i.e. they have some access to the propositions that constitute their evidence, first, and, second, the individual knows or justifiably believes with respect to any proposition whether or not it is part of their evidence. With respect to both layers, we can ask what the relevant sort of access is, but usually access internalists will identify the same kind of access as relevant. I will do the

---

[23] Take the case of an epistemically hapless individual that Pryor describes (2001: 114). The hapless individual is taught bad epistemic standards, for example they are not taught the difference between the likelihood of false negatives and false positives in assessing tests statistically. Additionally, they lack the intellectual capacities to discern the defects of what they have been taught themselves. It seems plausible that we would not *blame* this individual for having false statistical beliefs but we would nevertheless *criticise* them (unless their intellectual capacities are so reduced that we no longer treat them as rational subjects like ourselves, see Pettit and Smith 1996). In some sense, epistemic criticism, in virtue of the fact that it is agent-directed, is also a form of blame: we blame someone for having unsupported beliefs. But, intuitively, it seems that blame, even purely epistemic blame, is a more holistic evaluation of someone's beliefs where we take into account mitigating circumstances of all kinds, like intellectual capacities and education. That is why I set epistemic blame apart from epistemic criticism, and avoid speaking of deontology.

[24] Many abilities we have do not require the subject to have voluntary control over exercising the ability, see Chuard and Southwood 2009.

same. Importantly though, it is the requirement of access to whether or not some proposition is part of one's evidence, that turns a view into an access internalist justification theory. Even certain epistemic externalist views believe that individuals always know what their evidence is. So, access internalists suggest that it is in virtue of the fact that individuals are in a position to know with respect to any proposition whether or not it is part of their evidence that we can ask of them to believe what they should believe and to refrain from believing what they ought not believe.

Let us start with the access to the evidence itself: what is the evidence that is available to an individual in a context? Norman's case, given the non-inferential reading, and Barry's case, I said, suggest that the available evidence is whatever is fairly directly accessible to an individual in a context: the things they can see, the inferences they can reasonably be expected to draw etc. This is also what Alston suggests: the common-sense notion of accessibility that is linked to justification is something along the lines of "fairly direct accessibility" (1988: 275). The reasoning behind Alston's suggestion is the following. We do not think that individuals ought to engage in lengthy research or investigation before they can have justification to believe something. You can justifiably believe that your bike is in the shed even if you have not checked on it in a couple of days. On the other hand, we seem to think that individuals cannot miss the obvious and be justified (see Gibbons 2006). If you cannot find your sunglasses, even though you have been looking for them everywhere in your flat, you will not be justified in believing that you lost them when you forgot that they are sitting in your hair. Or, if you fail to put two and two together, as one says, because you have seen Odette and Owen together a lot, and Odette told you about her boyfriend Owen, you do not have justification to believe that Odette's boyfriend is another Owen.

Alston admits that he does not know how to make the notion of fairly direct accessibility more precise. Gibbons, in turn, suggests that the common-sense notion of accessibility is the notion of what we are in a position to know in a context without significantly changing our epistemic situation (2006: 28). What we are in a position to know differs from what we are capable of knowing because we are able to know all sorts of things with enough time and

resources that we are not already in a position to know. I am, for example, in a position to know the results of some major scientific finding at some point in the future when it hits the news because I will be alive, able to read the news and comprehend what is reported etc. But I am not now in a position to know the finding with the resources available to me now. The things that I can know without significantly changing my epistemic situation include, amongst many others, certain facts of my immediate environment (e.g. that the window is open), some sensations (e.g. that I am slightly hungry), some memories (e.g. that Williamson argues that Audi is wrong in one of his papers), some conclusions of inferences I could fairly easily draw (that the coffee that I made earlier is cold now). Those are things I am in a position to know without having to change my epistemic situation significantly. I will say that these are the things that I am in a position to know *readily*:

> *Ready access*: S has ready access to the fact that *p* at t if and only if S is in a position to know that *p* without significantly changing their epistemic situation at t.

Other labels for "ready access" are: that which we can know "on the spot", or that which is "unproblematically given" (Kelly 2008: 943), or that which we can know without further research or investigation.[25]

How are epistemic situations defined? Here is what Gibbons suggests: epistemic situations are defined epistemically (2006: 34). The following, he says, is what should hold true about epistemic situations in any event: If *S* has justification for believing that *p* in epistemic situation *E* and *S** does not have justification for believing that *p* in *E**, then *E* is not the same epistemic situation as *E**. It is likely that in many cases, our judgment of whether someone has justification for believing that *p* in different situations will reveal whether they are in the same epistemic situation rather than the other way around. What this means is that it is likely that we will have to rely on

[25] I do not mention "direct access" because direct access has sometimes been identified with non-inferential access (see Fumerton 1995: 83) but, surely, some more or less straightforward inferences are readily accessible to us.

intuitive judgments about whether someone has justification for a belief in a context to resolve the necessary vagueness that the notion of the readily accessible carries.

We saw that a number of methods may contribute to making a fact readily accessible: perception, introspection, memory, inference, amongst others. The readily accessible facts will furthermore lie on a spectrum of more or less direct accessibility: certain things are particularly easy to see, others can be spotted if one looks properly; some inferences are almost immediately drawn, others require a bit of reflection. What emerges is a sphere of readily accessible facts the boundaries of which will sometimes be vague, but also often clear: lengthy reflection, years of therapy, focused observation but also sobering up or reducing the number of simultaneous tasks are all ways to significantly change one's epistemic situation. Intuitive judgments about what someone knows and what they have justification to believe in a context are likely to be helpful in drawing that boundary in a specific context because we are likely to have more reliable intuitions about knowledge and justification than about relative accessibility of evidence.[26]

One will have noted that ready access is not restricted to mental states, but includes facts about one's surroundings. The claim that an individual's evidence is that which is "unproblematically" given to the individual has often been interpreted as an argument for evidential internalism, the view that an individual's evidence is constituted by their purely internal mental states (see, e.g., Audi 2001; Huemer 2007; Smithies 2019). Clearly, this line of reasoning relies on the intermediate assumption that only purely internal mental states are unproblematically given to an individual. As far as I can see, this is not supported by an intuitive notion of what is unproblematically accessible to an individual: if you are a normally developed human adult, and not asleep or in a coma, you have ready access to the fact that you stained your shirt during lunch, for example. As we will see in a moment, the focus of many access internalists on purely internal states derives not from a common-sense notion

[26] Similarly, Williamson argues in his account of knowledge that we will have to rely on our intuitive knowledge attributions to see whether the conditions for knowledge are satisfied in a specific case, because we more likely to have reliable intuitions about knowledge itself then about its satisfaction conditions (2000: Chap.4).

of access but specifically from considerations about what individuals have justification to believe in bad cases. Ready access, in contrast, is not limited to mental states but describes a sphere of readily accessible facts within *and* around the individual.

So, the first access claim is an access-based view on evidence: access internalists may plausibly hold that an individual's evidence is the totality of propositions they know readily.

The second and crucial access claim concerns access to the evidence itself. Access internalists claim in a second step that individuals can tell whether or not a fact is readily accessible to them in virtue of the fact that it is readily accessible to them. Take Norman and Barry in their respective epistemic situation. Norman has ready access to the fact that the proposition that the president is in NYC is part of his evidence because he has ready access to the fact that he 'clairvoyants' it. So he should believe that the president is in NYC and he would make a mistake if he did not believe it. Barry, in his deprived epistemic situation, has ready access to the fact that things seem to him a certain way, so he should believe that things are that way. Each in their respective epistemic situation responds well to the facts to which they have ready access, so they both have justification for their beliefs.

In many cases, the way in which individuals will know that a certain proposition is part of their evidence will be by virtue of self-knowledge: because they know that they see, or seem to see, or remember or seem to remember or infer that something is the case. Importantly, they know these things readily as well: they readily know that they see or seem to see that something is the case.

Assuming evidentialism, a possible version of access internalism is the combination of two access claims, based on the permissive notion of ready access:

(i) S always readily knows what S's evidence is, i.e. S readily knows the propositions that constitute their evidence;[27]

(ii) S always readily knows with respect to any proposition *p* whether or not *p* is part of S's evidence (see, e.g., Alston 1988, Gibbons 2006).

We may specify the access in the definition of access internalism in §1.1.2 as follows:

Ready access internalism: Whether or not S has justification to believe that *p* supervenes on facts to which S has ready access (see, e.g., Alston 1989; Gibbons 2006).[28]

After chapter 2, I will be exclusively concerned with ready access internalism.

Access internalists hold that when we criticise people for their beliefs, we criticise them by reference to the sphere of those facts that are readily accessible to them in a context (see Alston 1989; Gibbons 2006). If some fact is readily accessible to an individual but they missed it, they made a mistake and do not have justification to believe some *p*; if some fact is not readily accessible to them and they missed it, they did not make a mistake and may have justification to believe that *p* (depending on whether they believe what they should believe given the facts that are readily accessible to them); if they believe what they should believe given the facts that are readily accessible to them, their belief that *p* is justified.

1.3.1.4 Epistemic criticism and ready access to status

Some access internalists have argued that the way in which we criticise each other for our beliefs suggests that the notion of justification is not only or not

---

[27] Note that some externalist theories of justification endorse (i). Williamson's E=K, for example, endorses a slightly stronger version (i) according to which an individual's evidence is the totality of propositions they know.

[28] There are weaker versions of access internalism that only require that one typically or normally has access to the facts on which justification supervenes. These views will not matter in the following (for discussion see Alston 1989).

most importantly linked to ready access to facts about what our evidence is. They suggest that, in order to appropriately react to epistemic criticism, individuals must be capable of knowing (or justifiably believing) whether or not they have justification for a certain proposition. Here is Alston's summary of Ginet's argument to this effect:

> "(1) S ought to withhold belief that *p* if he lacks justification for *p*.
> (2) What S ought to do S can do.
> (3) Therefore, S can withhold belief wherever S lacks justification.
> (4) *S has this capacity only if S can tell, with respect to any proposed belief, whether or not S has justification for it*.
> (5) S can always tell this only if justification is always directly recognizable.
> (6) Therefore justification is always directly recognizable",
> (Alston 1986: 217; my emphasis).

According to this line of reasoning, individuals are generally capable to readily know with respect to any belief whether or not they have justification for having it. Note that this is more demanding than to know whether or not some proposition is part of your evidence because it may require complex reasoning to assess whether or not the evidence you have supports that belief. This may not imply that individuals need to be capable of explicitly formulating the epistemic standards according to which the belief is justified or not. But they must at least have an implicit understanding of when a belief is and when it is not held in accordance to good epistemic standards; maybe in a way in which native speakers can intuitively judge whether a sentence in their language is well-formed. Some access internalists have concluded from that that the notion of justification is linked to ready access to the justificatory status of one's beliefs itself (Chisholm 1977; BonJour 1985).[29]

[29] Contemporary forms of this view restrict it to positive cases: individuals are always in a position to know that they are justified to believe that *p* if they are but are not always in a position to know that they lack justification if they do (see McDowell 2011; Pritchard 2012; Greco 2014; Das and Salow 2016).

I will not be concerned with this strong access internalist view in the following chapters.

Note that Norman is likely to lack ready access to the justificatory status of his clairvoyant beliefs even on the non-inferential reading (at least in a context of 'clairvoyanting' for the first time). Norman does not know whether his clairvoyant beliefs conform to good epistemic standards, or even, assuming a more realistic setting than BonJour's own case, Norman has reasons to believe that his clairvoyant beliefs do not conform to good standards. He would have to collect enough inductive evidence about his own track-record before he would be in a position to know the justificatory status of his clairvoyant beliefs. Barry in the bad case, in contrast, is in a position to know that most of his beliefs are formed impeccably. Barry's issue is just that he is fed misleading evidence.

1.3.2 The Equal Justification Thesis: The NED Problem, 2nd take

Let us come back to the demon-world. So far I have considered what epistemic internalists say about the difference between Barry and Warren. But, importantly, epistemic internalists draw another lesson from the New Evil Demon problem. Barry, internalists suggest, is not only more justified than Warren but also *as justified as* Gary. Gary, recall, is Barry's physical and phenomenal duplicate who lives in the good case: if things appear a certain way in Gary's world, they are that way. So, in most cases, if Gary believes that *p*, he knows that *p*. The thesis that Barry is as justified in believing the same propositions as Gary is what I will call the "*Equal Justification Thesis*":

> *Equal Justification Thesis:* Take internal duplicates, G and B. G lives in the good case, B lives in the bad case. G and B have justification to believe the same propositions to the same extent.

Epistemic internalists claim that the Equal Justification Thesis is intuitively extremely plausible (see, e.g., Neta and Pritchard 2007; Littlejohn

2009). They further believe that epistemic internalism provides the best explanation for why it is true, if it is (Conee and Feldman 2001; Audi 2001; Wedgwood 2002; Huemer 2006; Smithies 2019).

The reasoning invoked to support the Equal Justification Thesis is related to the propriety of epistemic criticism: not only would Barry seem to conduct his epistemic affairs better than Warren, he would also not seem to conduct them worse than Gary. We are supposing that Gary and Barry are epistemic duplicates in the following sense: they believe exactly the same propositions, they undergo indiscriminable experiences, Barry seems to see and remember exactly the same things that Gary sees and remembers, they find the same things intuitive and are disposed to reason in precisely the same way. Barry is just unlucky that the evidence available to him is misleading in a way that Gary's evidence is not. According to epistemic internalists, it is difficult to shake the intuition that Barry and Gary are justification duplicates as well: while Barry does not know anything about the world around him, he has justification to believe exactly the same things as Gary.

Barry's case is what we may call a "*really bad case*": Barry has always lived in the demon-world and has always been fed really bad evidence insofar as the knowledge-conduciveness of the evidence is concerned. More realistically, we sometimes find ourselves in *"local bad cases"*. In local bad cases, little of the environment appears other than it is. Maybe you are dreaming extremely realistically and, in the dream, you reason properly. The epistemic internalist reasoning extends from cases like Barry's to such less drastic cases. According to epistemic internalists, if you respond carefully to the facts accessible to you in the dream, you will have justification to believe certain things even though your beliefs could not be further away from how things actually are. Whether globally or locally deceived, if the individual in a bad case is responding rationally to the facts to which they have access just as their duplicate in the good case would, the Equal Justification Thesis says that they have justification to believe the same propositions to the same extent.

Access internalists and mentalists give different explanations of the Equal Justification Thesis.

Gary and Barry have justification to believe the same propositions if they have the same evidence and have access to the same facts about their evidence according to access internalists. But, if ready access is the access relevant to access internalism, Gary and Barry do not have the same evidence, nor do they have access to the same facts about their evidence. As discussed, individuals have ready access to certain facts about their immediate environment, amongst others. But, then, Gary may have ready access to the fact that a window is open upon entering a room whereas Barry does not have ready access to the fact that a window is open upon seemingly entering a room because there is no window or room in the demon-world. Ready access is badly positioned for supporting an access internalist rationale of the Equal Justification Thesis.

However, access internalists have noted, while Barry does not have ready access to the fact that the window is open, he does have ready access to the fact that it seems to him that the window is open. And when a window is actually open in Gary's flat in the good case, it will also seem to Gary that the window is open in virtue of the fact that the window is open. In other words, Gary and Barry both have ready access to how things seem or appear to them and they appear the same to them. If the access relevant to the access internalist conception of evidence is not ready access but the specific form of ready access that Gary and Barry have to how things appear to them, they may have the same evidence and may have access to the same facts about their evidence. If that is the case, they would have justification to believe the same propositions to the same extent according to access internalism. Note, however, that giving up ready access *tout court* as the access relevant to justification immediately means to exclude certain aspects of reality that are intuitively relevant to justification from being relevant in virtue of the fact that we know about them empirically.

Seemings or appearances are mental phenomena that are directly accessible only to the individual who experiences them. The way in which we know any of our minds would seem to be special: we know our minds in a way that no one else does and in a way we do not know anybody else's mind or our surroundings. For example, I am capable to know that I am envious that

my friend wrote a novel or that I am moved by a book in a way that other people are not capable of knowing these things about me. It would seem that we have a *peculiar* way of knowing our own minds. Access internalists have characterised the peculiar kind of access individuals have to their minds in different ways: generally as "self-knowledge", "rational reflection" or "introspection" that gives "*a priori*", "direct", "special", "non-observational" or "non-empirical" access to one's mind (see, e.g., Ayer 1959; Alston 1986; Boghossian 1997; Audi 1998; Bernecker and Dretske 2000; Pryor 2001). This peculiar way of knowing one's mind has further been assumed to carry a certain kind of epistemic privilege that individuals would have with respect to their mind: the deliverances of our faculty for self-knowledge have been characterised as infallible, incorrigible or indubitable (Descartes ([1641]/1998; Chisholm 1977; Audi 2001). Even those who are doubtful that the deliverances of any faculty could be so distinguished agree that knowledge of our minds is *privileged* in the sense that beliefs about our own mind tend to be known more often than our beliefs about other people's mental states or our surroundings (see Byrne 2018: 5).

I will say that someone has *special access* to a fact when they have a peculiar and privileged way of knowing that fact:

> *Special access*: S has special access to the fact that *p* if and only if S knows that *p* in a peculiar and privileged way.[30]

Virtually any theorist about self-knowledge agrees that we have special access to some of our mental states.[31]

Access internalists claim that Gary and Barry have justification to believe the same propositions to the same extent because they have special access to the same states, such as their seemings and appearances, and to the fact that they have special access to these states.

---

[30] I chose the label "special access" over "introspective access" because the term "introspection" suggests a perceptual model of self-knowledge that many self-knowledge theorists reject (see Byrne 2018: Chap.1-3).

[31] Ryle (1949) and other radical behaviourists are an exception as they reject peculiar access (and accept privilege access only insofar as, by necessity, we collect the most extensive body of observations about ourselves).

Importantly, access internalists generalise from the (supposed) fact that, in the really bad case, Barry would seem to have justification to believe the same propositions to the same extent as Gary in the good case to local bad cases as well as to cases that are just located within the good case. For it would seem that, if we judge that Barry has justification for many of his empirical beliefs in virtue of his specially accessible states even though his epistemic situation is so very bad with respect to being accurate, then tracking the states that are specially accessible to an individual is likely to track what they have justification to believe in *any* epistemic circumstances. Consequently, access internalists suggest that if Gary in the good case believes that the window is open, his belief is justified in virtue of the fact that it seems to him that the window is open. In other words, it is in virtue of generalising from the (supposed) justification in the bad case to other cases that access internalists have largely settled on the claim that the access relevant to justification is special access (for discussion of the generalisation, see also Williamson 2000: Chap. 8). Their view can be defined as follows:

> *Special access internalism*: Whether or not S has justification to believe that *p* supervenes on facts to which S has special access (see, e.g., Audi 2001; Pryor 2001; Huemer 2007).

As mentioned, in chapter 2, I will show that hybrid access internalists cannot vindicate the Equal Justification Thesis (in its original form). But it is the Equal Justification Thesis only that would justify a focus on special access. As a result of chapter 2, I will suggest that hybrid access internalists should focus on the form of access that is more plausibly linked to justification, namely ready access. The discussion of chapter 3 and 4 will work with formulations of internalist positions in both internalism/externalism debates in terms of ready access.

Unlike the access motivation for epistemic internalism, the Equal Justification Thesis does not just apply to access internalists. Most mentalists also wish to vindicate the Equal Justification Thesis (Conee & Feldman 2001). Mentalists suggest that Gary and Barry have justification to believe the same

propositions to the same extent because they are in the same mental states that sufficiently support their beliefs. If the goings-on "inside" of Gary and Barry's mental systems are the same, they have justification to believe the same propositions to the same extent according to mentalists, regardless of the extreme differences between their surroundings.

1.4 Hybrid Epistemic Internalism: Challenges

The two main challenges for hybrid epistemic internalism derive from the two main motivations for epistemic internalism. The *access problem* is the issue that it would seem that, in virtue of the fact that external mental states are partly individuated by external factors, individuals do not have ready access to their external mental states in certain circumstances. If individuals do not have ready access to their external mental states, this may be an issue for hybrid access internalist theories (**§1.4.1**). The *equal justification problem* is the issue that, assuming state externalism, neither the preferred access internalist nor the preferred mentalist explanation for the Equal Justification Thesis is available (**§1.4.2**). The two challenges create different kinds of problems for hybrid epistemic internalism: while the equal justification problem undercuts one motivation for epistemic internalism, the access problem would result in a direct inconsistency between state externalism and access internalist views. The following chapters of this thesis are dedicated to the discussion of these challenges for hybrid epistemic internalism. I will close by sketching the plan for this thesis (**§1.4.3**).

1.4.1 The access problem

Many access internalists doubt that individuals have the kind of access that is relevant to justification to their external mental states. BonJour, for instance, writes:

> "The adoption of an externalist account of mental content would seem to support an externalist account of justification in the following way: if part or all of the content of a belief is inaccessible to the believer, then both the justifying status of other beliefs in relation to that content and the status of that content as justifying further beliefs will be similarly inaccessible, thus contradicting the internalist requirement for justification" (BonJour 1992: 136).

Let us call this the "access problem" for hybrid access internalism. We can state the access problem for hybrid access internalism as follows:

(i) If state externalism is correct, individuals lack a certain kind of access to their external mental states;
(ii) the kind of access individuals lack to their external mental states if state externalism is correct, is the kind of access relevant to access internalist theories of justification;
(iii) if individuals lack the relevant kind of access to their external mental states, this conflicts with the access requirements of access internalist views.
(iv) So, state externalism conflicts with the access requirements of access internalist views.

Traditionally, the access problem has been thought to come up with respect to an individual's *special access* to their mental states: the "relevant kind of access" in premise (i) - (iii) is filled in by "special access".

State externalists and internalists alike have been worried about whether state externalism is compatible with special access to external mental states ever since externalism about thought content was first proposed (see, e.g. Burge 1988; Boghossian 1989; Nuccetelli 1993; Falvey and Owens 1994; Brown 1995; McLaughlin and Tye 1998; Farkas 2008; Goldberg 2015). This is because two consequences would seem to follow for the epistemology of our mental states if some of our mental states are partly individuated by external factors. First, at least in certain circumstances, we would have to find out more about our environment in order to know what mental states we are in. This

contradicts the peculiarity of the access we would seem to have to our mental states. Second, we may make mistakes about our mental states that are due, not to any cognitive failings on our side, but to some ignorance about our surroundings. This contradicts the privilege of the access we would have to our mental states.

Further, in virtue of the Equal Justification Thesis, access internalists have generally considered special access to be the kind of access relevant to access internalism. It is therefore unsurprising that state externalism has often been discussed as a challenge to access internalism (see, e.g., Boghossian 1989; Chase 2001; Brueckner 2002; Pritchard and Kallestrup 2004; Brown 2007; Williamson 2007; Madison 2009 and Morvarid 2019).[32]

Since I believe that hybrid access internalist views should focus on ready access instead of special access, it will have to be seen whether the access problem comes up with respect to an individual's *ready* access to their external mental states as well. Note that one may think that state externalism does not lead to an access problem for hybrid access internalism if it is ready access that is at issue. For, even if individuals have to rely on empirical or inferential methods in order to know that they are in a particular external state, they may have ready access to their external states as long as the empirical or inferential evidence is readily accessible to them.

In chapter 3, I will argue that not only does a very similar access problem come up with respect to ready access to one's external states but we get an even better understanding of the workings of the access problem of state externalism if we focus on ready access instead of special access. As we will see, two consequences about the epistemology of our mental states that conflict with ready access to our mental states will follow from the truth of state externalism. First, at least in certain circumstances, we will have to find out *more* about our environment in order to know that we are in a particular mental state: what we readily know in a context is insufficient to know that we are in a particular external state. Second, we may make mistakes about our mental states that are due to being empirically ignorant about those parts of

[32] The majority of these thinkers is concerned with the consistency of *content* externalism with access internalism since other forms of externalism about mind have been neglected.

our surroundings that are not readily accessible to us. The access problem is pertinent if we replace the "relevant sort of access" in premise (i) and (iii) by "ready access".

Note that hybrid mentalism, in virtue of the fact that it does not formulate an access requirement on justification, does not confront the access problem. This has resulted in a state of the debate on the consistency of hybrid epistemic internalism where the consistency of state externalism and access internalism is habitually questioned whereas it is often assumed for mentalism, at least for externalism about thought contents (Conee 2007; Wedgwood 2002, 2017; Schoenfield 2015; as mentioned, the latter two do not think that mentalism is compatible with externalism about attitudes).

Note that if individuals lack ready access to their external states, this establishes premise (i) of the access problem for hybrid epistemic internalism. But premise (iii) does not have to be uncontroversial. Premise (iii), recall, says that it conflicts with the access requirements of access internalist views if individuals lack the relevant sort of access to their external states. But it is possible that, in those cases in which individuals lack ready access to some of their external states, an individual's justification would supervene on those states to which they still have ready access in that context, maybe relying on other forms of ready access. More generally, it is not immediately obvious what hybrid access internalists have to say about the justification for their beliefs that individuals have in those circumstances in which they lack ready access to some of their external states. In chapter 3 and 4, I will focus on the discussion of premise (i), i.e. whether state externalism undermines ready access to one's states. Although it is not exactly accurate, I will therefore also speak of the "access problem" for hybrid epistemic internalism when I mean the problem of whether we lack ready access to our external states if state externalism is correct. But, it will be interesting to inquire into premise (iii) in future research for it has, in comparison, been neglected.[33]

[33] Exceptions are Chase (2001), Brueckner (2002), Vahid (2003a), Brown (2007), and Morvarid (2019).

1.4.2 The equal justification problem

The *equal justification problem* is the issue that, assuming state externalism, neither the preferred access internalist nor the preferred mentalist explanation for the Equal Justification Thesis would seem to be available.

Let us start with hybrid access internalism. We saw that access internalists identify special access as the relevant sort of access for justification on the basis of the Equal Justification Thesis. Let us assume, then, that the relevant sort of access is special access. Do Gary and Barry have special access to the same states if state externalism is assumed? It would not seem that they do. Barry lives in a really bad case: he has always lived in the demon-world and has never interacted with real things and real people. Assuming the different versions of state externalism, many of the propositions that Barry believes will have a different content from the propositions that Gary believes in indiscriminable circumstances: Barry does not have beliefs about water but about demon-water; Barry does not have beliefs about arthritis, joints or muscles but about the demon versions of those; if Barry thinks that that "apple" is overripe, he does not have the same belief as Gary who thinks that that (real) apple (a) is overripe. But if Barry does not have thoughts with the same contents as Gary, it is likely that he will have special access to other thought contents than Gary. If Barry has special access to other thought contents than Gary, it is likely that he has special access to other facts about his evidence and has justification to believe different propositions from Gary. But this contradicts the Equal Justification Thesis. Call this the "*equal justification problem*" for hybrid access internalism.

Hybrid mentalists confront the same issue, just not mediated via the question of access. Consider external attitudes. Certain attitudes are not available to Barry in the really bad case because these attitudes require the world to be a certain way: Barry will not, for example, remember many things he believes he remembers. But Gary will remember those things. If Barry in the bad case cannot have certain attitudes towards some contents that Gary has towards the same (or not the same) contents, he will not be in exactly the same mental states as Gary. But then Barry it is likely that Barry has justification to

believe different things than Gary according to mentalism. But this contradicts the Equal Justification Thesis. This the "*equal justification problem*" for hybrid mentalism.

1.4.3 Plan for the thesis

This thesis has three further chapters.

In the next **chapter 2**, I will look into whether the equal justification problem can be solved for hybrid epistemic internalism. I will argue that even if we assume that the access relevant for justification is special access, hybrid access internalism cannot vindicate the Equal Justification Thesis in its original form. In the chapters 3 and 4, I will therefore focus on statements of internalist positions in terms of ready access. However, there is a much weaker variant of the Equal Justification Thesis — the Indiscriminable Justification Thesis — that would seem to be entailed by hybrid access internalism as well as by hybrid mentalism. Whether it is entailed depends on results of the discussion in chapter 3 and 4. In the very last part of this thesis (§4.4), I will suggest that the Indiscriminable Justification Thesis solves the equal justification problem for hybrid epistemic internalism.

Chapter 3 will address the access problem for hybrid access internalism (or, more precisely, the first premise of it). I will apply the widely-known "*discrimination argument*" against *special* access to external mental states to *ready* access to one's external mental states. I will argue that the discrimination argument against ready access to one's external states is sound. Furthermore, the circumstances in which we lack ready access to our external mental states are sometimes relevant to us. The result of chapter 3 is that, if state externalism is true, we sometimes do not know that we are in a particular external state rather than some relevant alternative external state unless we significantly change our epistemic situation. Premise (i) of the above access problem for hybrid access internalism is true.

In the last **chapter 4**, I will argue that it is the defining feature of state externalism to undermine an individual's ready access to their external states

in certain circumstances. More precisely, I will argue that state externalism is defined as the position according to which “mental state switching” is possible. Mental state switching is a way for an individual’s mental states to change (or “switch”) in a way that the individual is unable to register purely out of empirical ignorance. In other words, I will defend an epistemic definition of the internalism/externalism debate about mind against metaphysical competitors. I will close by suggesting that this results in an unified, epistemic definition of the internalism/externalism debates in the philosophy of mind and epistemology.

Chapter 2

2 - The Equal Justification Problem for Hybrid Epistemic Internalism

2.1 The Equal Justification Thesis and State Externalism

It follows from epistemic internalism that if two individuals are the same "on the inside", they have justification to believe the same propositions to the same extent:

> *Epistemic internalism*: Internal duplicates have justification to believe the same propositions to the same extent.

Access internalists and mentalists, we have seen, disagree about when individuals are internal duplicates in the relevant sense. Individuals are internal duplicates in the mentalist sense if and only if they are in the same mental states. Individuals are internal duplicates in the access internalist sense if and only if they have special access to the same facts about their evidence.

One central motivation for epistemic internalism is the Equal Justification Thesis. Epistemic internalists claim that it is intuitively extremely plausible that an individual in the bad case, I called him Barry, has justification to believe the same propositions to the same extent as his epistemic duplicate, that was Gary, in the good case:

> *Equal Justification Thesis:* Take physical and phenomenal duplicates G and B. G lives in the good case, B lives in the bad case. G and B have justification to believe the same propositions to the same extent.[34]

According to epistemic internalists, the best explanation for why the Equal Justification Thesis is true is that Gary and Barry are internal duplicates in the epistemic internalist sense as well. According to access internalists, Gary and Barry have justification to believe the same propositions to the same extent because they have special access to the same facts about their evidence.

[34] Of course, some form of internalism about phenomenal properties is assumed here.

According to mentalists, Gary and Barry have justification to believe the same propositions to the same extent because they are in the same mental states.

The argument that links epistemic internalism and the Equal Justification Thesis is straightforward:

(i) Internal duplicates have justification to believe the same propositions to the same extent (Epistemic internalism).
(ii) Gary in the good case and Barry in the bad case are internal duplicates.
(iii) So, Gary and Barry have justification to believe the same propositions to the same extent (Equal Justification Thesis).

The issue is that if state externalism is correct, premise (ii) is false: Gary and Barry are not internal duplicates, neither in the access internalist nor in the mentalist sense. But then neither hybrid access internalism nor hybrid mentalism entails the Equal Justification Thesis. That was the equal justification problem. This chapter is about whether there are any ways for hybrid epistemic internalism to solve the equal justification problem. The equal justification problem is solved if hybrid epistemic internalism entails the equal justification thesis, or a plausible variant. So, this chapter is about whether hybrid epistemic internalists can motivate their view by the Equal Justification Thesis (or a variant); it is not about whether the Equal Justification Thesis is true or plausible.

It is obvious that Gary and Barry are not internal duplicates in the mentalist sense. Recall, the most famous sceptical scenarios likes Descartes' demon scenario or Putnam's brains-in-vats scenario are not only bad cases but *really bad cases*: we imagine that Barry was born (by the demon, I guess) in the demon-world or has been created in the vat.[35] Many of their natural kind thoughts, of their socially-individuated thoughts and of their singular thoughts will have different contents. Further, even if we focus on states with the same contents, Gary will know some things because he lives in the good case

[35] In this chapter, I will assume that Barry lives in a vat rather than in the demon-world for conciseness.

whereas Barry will falsely believe some of the things that Gary knows. So they are not in the same factive states if attitude externalism is assumed. It is consistent with mentalism that Gary and Barry may nevertheless have justification to believe the same propositions to the same extent as it is consistent with mentalism that non-duplicates have justification to believe the same propositions. But they could not have justification to believe the same propositions *in virtue* of having the same mental states.

The same problem arises in a slightly less obvious way for access internalists. Most state externalists are *compatibilists*: they believe that individuals have special access to their external thought contents and factive states (see, e.g. Burge 1988; Heil 1988; Nuccetelli 1993; Falvey and Owens 1994; McLaughlin and Tye 1998; Goldberg 2015; Das and Salow 2016). We may define:

> *compatibilism*: individuals have special access to their externally-individuated mental states.

If compatibilism is correct, Gary and Barry do not have special access to the same mental states. If they have special access to different states, they will have special access to different facts about their evidence. For example, Gary will have special access to the fact that he can see that he has hands whereas Barry will not have special access to that fact. If Gary and Barry have special access to different facts about their evidence, they may not have justification to believe the same propositions to the same extent. It is consistent with access internalism that they may nevertheless have justification to believe the same propositions to the same extent as it is consistent with access internalism that non-duplicates may have justification to believe the same propositions to the same extent in some cases. But they could not have justification to believe the same propositions to the same extent *in virtue* of having special access to the same facts about their evidence.

I will discuss three ways in which hybrid epistemic internalists have tried or could try to solve the equal justification problem. In **§2.2**, I will look into ways in which access internalists and mentalists have tried to show that

premise (ii) is true despite state externalism: somehow, Gary and Barry are still internal duplicates. The other two attempts change the datum, i.e. they propose a new premise (i), and argue that epistemic internalism is in fact a weaker thesis. The weaker forms of epistemic internalism, and specifically hybrid epistemic internalism, they argue, entail weaker but more plausible versions of the Equal Justification Thesis. In **§2.3**, I will discuss the Counterpart Justification Thesis that states that counterparts like Gary and Barry have justification to believe *counterpart* propositions to the same extent. In **§2.4**, I will consider the Indiscriminable Justification Thesis that states that individuals with indiscriminable mental states, or as I will call them "indiscriminables", like Gary and Barry have justification to believe *indiscriminable* propositions to the same extent. I will argue that the Indiscriminable Justification Thesis is plausibly entailed by hybrid epistemic internalism but that its final assessment will have to wait until certain results are established in chapter 3 and 4. In the very last part of the thesis, I will suggest that a qualified version of the Indiscriminable Justification Thesis may solve the equal justification problem for hybrid epistemic internalism.

2.2 Vindicating the Equal Justification Thesis

In this section, I will discuss how access internalists (**§2.2.1**) and mentalists (**§2.2.2**) have tried to show that individuals in bad cases and good cases may be internal duplicates even if state externalism is assumed.

2.2.1 Access internalists go recent and incompatibilist

Earlier I discussed really bad cases where Barry has always lived in an environment where most of his beliefs are false. We can distinguish really bad cases from "*recent bad cases*": in a recent bad case, Barry has only recently been moved ("been" because Barry is ignorant of the move) to the bad case.

Neta and Pritchard, for example, identify the following as the actual intuition behind the Equal Justification Thesis: "[...] The extent to which S is justified at t in believing that p is just the same as the extent to which S's recently envatted duplicate is justified at t in believing that p" (Neta and Pritchard 2007: 381). The appeal of focussing on recent bad cases is obvious: according to externalism about natural kind thoughts and about socially-individuated thoughts, the thought contents of the recently deceived individual will be the same as the thought contents of their duplicate in the good case. Like this, some forms of content externalism are kept from coming into effect.

One problem with restricting the Equal Justification Thesis in this way is that it is doubtful that it would result in a total state overlap between Gary and Barry. Take demonstrative thoughts. Suppose that Gary believes that that apple is overripe. Assuming singular externalism, recently envatted Barry cannot have the same thought since he is not seeing an apple. Furthermore, Gary will probably have special access to those demonstrative thoughts. Gary may, for example, infer from his belief "this apple is overripe" to "I can see that this apple is overripe" (see Byrne 2012). So Gary can know some of his factive states by introspection. But Barry is lacking some of these factive states. If Gary has special access to his demonstrative thoughts and factive states, he will have special access to different states than Barry, and if he has special access to different states, then he will have special access to different facts about his evidence.

Another problem with restricting the claim to recently envatted individuals is that the restriction is badly motivated. As was discussed in §1.3.2, the Equal Justification Thesis is motivated by reflections on the propriety of epistemic criticism. Gary and Barry both respond perfectly rationally to certain facts available to them in a context where everything seems the same to them. Since any differences between their respective epistemic situations is beyond what they are in a position to know or justifiably believe, it is difficult to shake the intuitions that they have justification to believe the same things to the same extent. But these considerations apply to recent bad cases as much as to really bad cases. We could imagine an individual who first lives in the good case, and is then

moved to the vat-verse to stay. In the vat-verse, some of their states will immediately change, such as their demonstrative thoughts and some factive states. But they will use many of the same concepts that they used outside the vat-verse, until, after a while, their expressions adapt to the new environment according to content externalism. In the last stage, many of their thought contents and factive states will have changed. Assuming compatibilism, they will have special access to many of them which will affect to which facts about their evidence they have special access. But all of this may take place unbeknownst to the individual themselves.[36] If Barry in a recent bad case has justification to believe the same propositions as Gary in the good case because everything seems the same to him and he has the same epistemic dispositions than Barry in the really bad case has justification to believe the same propositions as Gary as well. Restricting the Equal Justification Thesis to individuals in recent bad cases is *ad hoc*.

That a limitation to recent bad cases would be *ad hoc* is further underlined if we compare Barry's situation in the really bad case relative to Gary's situation in the good case with Twin Oscar's situation on Twin Earth relative to Oscar's situation on Earth. Twin Oscar, as a competent believer in his environment, will have many justified beliefs about his environment. But, assuming the content externalist version of the Twin Earth story, it will not be the same beliefs that Oscar has justification to believe on Earth. Williamson writes:

> It is not in dispute that we can pick an example in which Oscar's belief that there are pools of water is justified. Perhaps he is swimming in one. Thus Oscar has the justified belief that there are pools of water. But Twin Oscar lacks a justified belief that there are pools of water, because he lacks the belief that there are pools of water. Thus Oscar and Twin Oscar differ in their justified beliefs, even though they are internal duplicates. Likewise, of course, Twin Oscar has the justified belief that there are pools of twater, while Oscar lacks a justified belief that there are pools of twater, because he lacks the belief that there are

[36] This is consistent with the compatibilist claim that at any specific moment, individuals are able to know that they are thinking a thought with a particular content (see, e.g. Burge 1988; Falvey and Owens 1994). But they will be ignorant of the fact that their thought contents changed from some non-vat-content to vat-content.

> pools of twater: that is just another difference in justified belief between Oscar and Twin Oscar (Williamson 2007: 108).

(When Williamson speaks of "internal duplicates" here, he means internal duplicates in the sense of duplicates of physical or phenomenal states. Oscar and Twin Oscar are not internal duplicates in the mentalist sense, of course: they are not in the same mental states given content externalism.) But Oscar's and Twin Oscar's situations seem, by hypothesis, exactly the same to them, and they respond perfectly analogously to their respective situations. The same considerations that support that Barry in the really bad case has justification to believe the same propositions as Gary in the good case support that Twin Oscar on Twin Earth has justification to believe the same propositions as Oscar on Earth (and *vice versa*).

In fact, Barry's and Gary's comparative situation may moreover be similar to Twin Oscar's comparative situation with Oscar in the following respect. Depending on one's views on the concepts that Barry acquires in the really bad case, it is not even clear that many of Barry's beliefs will be false in the really bad case. When Barry uses the expression "water" in the vat, for example, he will refer to what we may call "vater", which refers to the thing that he vat-drinks, vat-washes with, vat-swims in etc. in the vat-verse. Depending on what exactly the meaning of "water" is in Barry's language, it is unclear that most of his vater-beliefs will be false. Vater is of course very different from what Barry imagines it to be: it is not an organic liquid but instead is made out of electronic signals. But it is unclear that that would keep Barry from successfully referring to vater even though he has many wrong ideas about its underlying metaphysical nature (the ancient Greeks, say, successfully referred to water even though they had many wrong ideas about its metaphysical nature). If Barry's vater-beliefs are not false, his vater-beliefs compare to Gary's water-beliefs exactly like Twin Oscar's twin water-beliefs compare to Oscar's water-beliefs.

Bad cases and twin cases form a unity to which the same considerations apply that epistemic internalists like to stress in order to motivate the Equal Justification Thesis: everything would seem to seem the same to them and

they act exactly alike, most importantly epistemically. Recent bad cases cannot be separated from the lot without arbitrariness. Or at least, by excluding individuals in really bad cases as well as twin cases from the scope of their claims, access internalists fail to give a principled account of how to weigh the considerations that motivate the Equal Justification Thesis against the appeals of state externalism.

One could try to exploit the difference between propositional and doxastic justification and argue that Barry in the really bad case and Twin Oscar *have justification* to believe the same propositions as Gary and Oscar, respectively, even if they in fact believe different propositions (see Williamson 2007: 108-9, for discussion). But even though this proposal is likely to fail for a number of reasons (such as making justification for indefinitely many false propositions cheaply available), it is not an option open to hybrid access internalists. Barry and Twin Oscar do not have special access to facts about their evidence that would concern water. So access internalists would have no explanation to give for why Barry and Twin Oscar would have propositional justification for any beliefs about water.

A more natural next move for access internalists is to reject compatibilism and argue that individuals only have special access to their purely internal mental states. Gary and Barry, and Oscar and Twin Oscar while being in different externally-individuates states, are, by hypothesis, in the same purely internal states. The resulting view consists of two claims:

(i) justification supervenes on facts to which the individuals has special access;

(ii) Individuals only have special access to their purely internal states.

While such a view is coherent, it need not interest us in the present context.[37] The question at present is whether hybrid access internalism entails the Equal Justification Thesis. But recall how hybrid access internalism was defined:

[37] See, esp. Smithies (2019) for an argumentative project of this kind; but important influences and precursors are phenomenal intentionalists like Horgan and Tienson (2002) and Farkas (2008), and phenomenal conservatives like Huemer (2007).

> *Hybrid access internalism*: a form of access internalism according to which, whether an individual's belief that *p* is justified at t may be partly determined by the individual's external mental states at t.

The above view is not a hybrid access internalist view: an individual's external states do not partly determine whether the individual has justification for a certain belief. Instead, such a view is fake hybrid: while state externalism is accepted, it is sidelined from playing a part in the justification theory.

The Equal Justification Thesis in its original form was the reason why access internalists have considered special access to be the access relevant to their justification theory. But special access provides Gary and Barry with access to the same facts about their evidence only given substantial assumptions about which mental states they share and which states individuals have special access to. Neither of these assumptions is consistent with hybrid access internalism. Hybrid access internalists should focus on the more plausible notion of *ready* access in the formulation of their views.

Having special access to one's mental states is one way in which individuals have ready access to their states. But given that differences in the facts to which Gary and Barry have ready access will extent to the facts to which they have special access, assuming a widely-held compatibilism, nothing is achieved by excluding facts that are intuitively accessible to Gary and Barry from being part of their epistemic situation. I will assume ready access to be the pertinent form of access to hybrid access internalist views from here onwards.

2.2.2 Mentalists go recent and partial

For hybrid mentalists who are motivated by the Equal Justification Thesis, the issue presented by state externalism is painfully direct: Gary and Barry simply do not have the same total set of mental states if state externalism is correct. So they are not mental duplicates, and may not be equally justified. The tool

that mentalists have available to try and cut the pie in a way that Gary and Barry are internalist duplicates despite state externalism is cruder: exclusion.

Mentalists proceed in two steps to argue that Gary and Barry are internal duplicates while partially accepting state externalism:

(i) like access internalists, they stipulate that Barry has only recently been envatted or moved to the demon-world;
(ii) they argue that factive states are inapt to be epistemic reasons.

The combined view is that Gary and Barry are mental duplicates because they have the same non-factive states, some of which may have externally-individuated contents, which, it is supposed, are the same in virtue of (i) (see Wedgwood 2002; Conee 2007; Schoenfield 2015).

The limited effectiveness of (i) was just discussed: despite (i), some of Gary's and Barry's singular thoughts and factive states will differ. If that is the case, it is likely that Gary has justification to believe some things that Barry does not have justification to believe.

To take care of factive states, some mentalists argue that factive states cannot *doxastically* justify other states (see Wedgwood 2002, 2017; and Schoenfield 2015). What matters for doxastic justification, recall, is not only that the individual possesses sufficient evidence for their belief but that the evidence causes or sustains the belief in the right way. In discussing the consistency of hybrid epistemic internalism, we are concerned with propositional justification, as pointed out in §1.1.2. So even if factive states could not doxastically justify other beliefs, they may be able to *propositionally* justify other beliefs.

Let me nevertheless point out why the mentalists' argument is unlikely to be successful. In a nutshell, Wedgwood and Schoenfield's view is that only non-factive mental states could proximally cause a rational belief revision. I will focus on Wedgwood's argument.[38] Here is Wedgwood's basic framework. Wedgwood claims that rational belief revision is a matter of following what he

[38] Schoenfield explicitly relies on Wedgwood's argument in the defence of the crucial step in her argument (see Schoenfield 2015: 261).

calls “basic rules”. Basic rules are of the form “If C, then *phi*”. Importantly, basic rules cannot be analysed at the folk-psychological level into a series of more basic rules. Wedgwood further claims that we can infer the correct basic rules from “fully-articulated folk-psychological causal explanations” (2002: 357): the state that is mentioned in a fully-articulated folk-psychological explanation as the most proximate cause of one’s belief identifies the trigger condition in a basic rule.

Now consider the following example. Wedgwood suggests that one follows a rule such as “Add salt when the water is boiling” by following the basic rule “Add salt when you believe that the water is boiling”. This is because, Wedgwood claims, the proximate explanation of one’s attempt to add the salt is not the fact that the water is boiling, but rather one’s belief that the water is boiling. “Similarly”, Wedgwood continues, “perhaps the proximate explanation of one’s belief that the water is boiling is not the fact that the water is boiling, but one’s having an experience that represents the water as boiling” (2002: 356). If that is correct, only mental states could cause further mental states. But, crucially, Wedgwood believes that not all mental states could be the proximate cause of further mental states but only non-factive states could be proximally causally efficacious.

His argument to this effect rests on three assumptions. First, Wedgwood assumes that knowledge is a compound state: knowledge is partially constituted by belief (361). Second, he appeals to what he calls a plausible general principle about causal explanation involving states that are partially constituted by another state: “if one fact is partially constituted by a second, and a certain effect would still have been produced even if the second fact had obtained while the first fact had not, then if either fact explains that effect, it is the second fact rather than the first” (362). Third, he appeals to a further general principle about causal explanations (which is said to be an effect of the first), namely that there must be a certain sort of proportionality between the explanandum and the explanans: the explanans must be sufficient in the circumstances to produce the explanandum; but it also must not contain any irrelevant elements that could be stripped away without making it any less sufficient to produce the explanandum (363).

Here is Wedgwood's argument. For any case where a factive state would seem to figure in a folk-psychological explanation of a rational belief revision, it would seem that one could explain the belief revision just in terms of the compound non-factive state. Suppose that we want to explain why a thinker comes to believe that *q*, and could explain that either by appealing to their knowledge that *p* or their belief that *p*. Had the agent merely believed that *p*, and not known that *p*, they would still have come to believe that *q*, in exactly the same way, Wedgwood says (361). So, according to Wedgwood it is the thinker's *believing* that *p* and not their knowing that *p* that really explains their forming the belief that *q*.

Interestingly, according to Wedgwood, we need not worry that this argument would expand to exclude states with externally-individuated contents from appearing in basic rules in virtue of the proportionality worry. This is because, the issue with factive states, according to Wedgwood, is not that they depend on the environment in a way that makes it hard to account for their involvement in mental causation but "the trouble is that what one knows is too dependent on the environment to give a *suitably proportional explanation* of one's forming this belief" (363, my emphasis). So, according to Wedgwood, it is not an issue if a non-factive state with an externally-individuated content causes another non-factive state with an externally-individuated content: "If the content of the belief that figures in the explanandum is itself determined by the thinker's relations to her environment, it is only to be expected that the explanation of this belief will involve mental states whose content is also determined by the thinker's relations to her environment" (363).

Wedgwood's argument that only non-factive states with possibly externally-individuated contents can figure in basic rules is inconclusive. First, not everybody accepts that factive states are compound states: Williamson, famously, believes that factive states are as little compound states as non-factive states are (2000: Chap. 1). To those, Wedgwood's first causal explanation principle would simply not be relevant to explanations in which a factive state is the most obvious explanatorily relevant candidate state. Second, Wedgwood seems to ignore the fact that some rational belief revisions

will result in knowledge. But if the only issue with excluding factive states from figuring in basic rules is a proportionality worry, then it is "only to be expected" that the explanation of why someone came to know something may involve other things that the person knows. Say, I know that *p*. I may infer that I know that *p or q* even if *q* is a wild guess because I know that *p*. My knowledge that *p* will be the natural causally proximate candidate state that figures in a fully articulated and proportional folk-psychological causal explanation of my knowledge that *p or q*.

In sum, there does not seem to be a way to avoid having to make quite a few arbitrary cuts and stipulations in an attempt to turn Gary and Barry, or Oscar and Twin Oscar, into internal duplicates in the access internalist or in the mentalist sense if state externalism is accepted.

2.3 Vindicating the Counterpart Justification Thesis

In the last section, I looked into ways epistemic internalists have attempted to defend the consistency of the Equal Justification Thesis with state externalism, either by thwarting the effects of state externalism or by excluding it. In this section, I will look into an attempt to solve the equal justification problem that respects state externalism.

If state externalism comes into effect, Gary and Barry will be in a number of different mental states. Nevertheless, it seems that there is an important correspondence relation between Gary's and Barry's epistemic situations. Gary has justification to believe that seawater is not drinking water because he drank it and it was too salty, or because he heard that salt withdraws water from one's body etc. Barry has justification to believe that seavater is not vat drinking water because he vat-drank seavater and it was too vat-salty, or because he vat-heard that vat-salt vat-withdraws vater from one's vat-body etc.

The relation between Gary's and Barry's epistemic situations, it is suggested, is that of *counterpart situations*. Conee writes about Oscar and Twin Oscar, for example: "the water proposition must be as well justified for the Earthling as some counterpart of that proposition is for the internal duplicate Twin Earthling" (2007: 52). So, while state externalism is incompatible with the fact that individuals in really bad cases or in twin cases and individuals in good cases have justification to believe the same propositions to the same extent, they may have justification to believe counterpart propositions to the same extent.

The proposal corresponds to a general weakening of epistemic internalism to a claim not about internal duplicates, but about counterparts. Counterpart epistemic internalism entails a correspondingly weaker version of the Equal Justification Thesis:

(i) Counterparts have justification to believe counterpart propositions to the same extent (Counterpart epistemic internalism).
(ii) Gary in the good case and Barry in the bad case (Oscar and Twin Oscar) are counterparts.
(iii) So, Gary and Barry (Oscar and Twin Oscar) have justification to believe counterpart propositions to the same extent (Counterpart Justification Thesis).

So the following may be entailed by hybrid epistemic internalism even if the original Equal Justification Thesis is not:

> *Counterpart Justification Thesis*: Take the counterparts, G and B. G lives in the good case, B in the bad case. G and B have justification to believe *counterpart* propositions to the same extent.

As we will see, the main issue with the counterpart justification thesis is that it would not seem possible to define the counterpart relation in a way that is consistent with state externalism. This is an issue for those hybrid epistemic internalist views that are the subject of this chapter: epistemic

internalist views that wish to do two things simultaneously, namely attributing to external mental states a proper role in justifying relationships and motivate their views by virtue of the Equal Justification Thesis or a variant. It may of course not be a problem to other epistemic internalist views.

I will introduce how the Counterpart Justification Thesis has been defended by a number of theorists in **§2.3.1**. These views crucially appeal to the difference between enabling and justifying conditions where it is suggested that the former may be externally-individuated whereas the latter are internally determined. However, existing accounts identify justifying conditions generally with narrowly-individuated mental states and properties. So, like the access internalist views that were considered in the last section, they fail to propose real hybrid accounts of the Counterpart Justification Thesis. In **§2.3.2**, I will see whether a proper hybrid account of the Counterpart Justification Thesis can be defended. As we will see, both the separability of enabling and justifying conditions as well as the counterpart relation create more issues than counterpart epistemic internalists have assumed.

2.3.1 The Counterpart Justification Thesis

Existing defences of the Counterpart Justification Thesis come in three steps. The first step is to defend the separability of individuation and justification: the claim is that environment-dependent state individuation does not entail environment-dependent justification **(§2.3.1.1)**. The second step is to spell out the counterpart relation between counterpart mental states **(§2.3.1.2)**. In virtue of their environment-independence, counterpart internalists claim that the same justifying conditions may be present across distinct environments and justify counterpart states **(§2.3.1.3)**.

2.3.1.1 Separability of individuation and justification

The key distinction that drives the claim that states may be individuated by the environment and yet be justified to the same extent is the distinction between justifying and enabling conditions (Audi 2001; Conee 2007; Madison 2009). Justifying conditions are those conditions that determine whether someone is justified in holding some belief, for example whether they apply good epistemic standards and respect the evidence they have. Think back to Warren who also lives in the demon-world but tends to disregard his evidence. Warren lacks reasons for believing certain things (or believes it for the wrong ones) and therefore is not as justified as Barry in his beliefs. We can also say that Warren does not respond to certain *justifying conditions* that Barry responds to.

The situation between Gary and Barry is a very different one. The reason why Barry does not have justification to believe the same propositions as Gary is not that he does not satisfy certain justifying conditions, but that he cannot grasp the propositions that Gary believes (and *vice versa*). Take Gary's belief that seawater is not drinking water. Barry does not have justification to believe that seawater is not drinking water not because he ignores some piece of evidence, but because he lacks the concept WATER. We can also say that Barry does not satisfy certain *enabling conditions* that Gary satisfies (and *vice versa*).

Clearly, a difference in justifying conditions is inconsistent with sameness of justification: justification directly supervenes on justifying conditions. If justification directly supervenes on justifying conditions, epistemic internalism is committed to the claim that justifying conditions are internal conditions in some relevant sense (spelled out differently by access internalists and mentalists). But since justifying conditions are claimed to be distinct from enabling conditions, epistemic internalists may not be committed to the claim that enabling conditions are internal conditions too. Audi, for example, suggests that

> "the *content* of my obligation can be external while its grounds are internal, just as the content of my belief about the water in my glass can be external though the grounds of that belief are internal" (2001: 37–38).

Madison claims:

> "The environment may determine which content a subject believes, but it does not determine which content the subject is justified in believing, any more than the fact that experience plays an enabling role in acquiring the relevant concepts fails to undermine the fact that our justification for the claim that 'all bachelors are unmarried men' is *a priori*" (2009: 180).

Defending counterpart mentalism, Conee stresses in a similar vein:

> "Mentalism is inclusive, but it is not noncommittal. Its strong supervenience thesis does exclude contingent or purely environmental variations from affecting justification" (2007: 59).

What these quotes suggest is the following picture. Epistemic internalism is committed to the internality of justifying conditions. Internality, it is suggested, is environment-independence: the justification of some state can remain constant through distinct environments as long as the internal justifying conditions stay the same. But the internality of justifying conditions is consistent, it is claimed, with the externality (i.e. environment-dependence) of enabling conditions.

2.3.1.2 Individuation: counterpart states

While Gary and Barry are in distinct states if state externalism is correct, there seems to be an important correspondence between Gary's and Barry's mental states. For example, if one travelled to the other in his environment for a sufficient period of time, given their general belief-forming dispositions, they would form exactly the same beliefs in any specific circumstance (and *vice versa*). Both Madison and Conee draw on the idea of *counterpart propositions* to specify what the correspondence consists in: while Gary and Barry do not

believe the same propositions if state externalism is assumed, they believe counterpart propositions, they suggest.[39]

Conee defines that "a counterpart to some proposition that is being considered is a proposition that differs from the considered one in an environmentally determined way, if at all" (Conee 2007: 52). Madison suggests something along the same lines: counterpart propositions are those propositions that "are as similar as possible while still being non-identical, as well as play an identical role in the subject's noetic structure" (Madison 2009: 182).

We can extract two claims by which counterpart propositions are specified: (i) they are distinguished by environmental factors only; and (ii) they share a functional role in the respective individuals' psychologies (this is how I understand "a subject's noetic structure").

We can extrapolate that Conee and Madison would specify counterpart attitudes in a similar way: a counterpart to some considered attitude is an attitude that differs from the considered one in an environmentally determined way, if at all; or, counterpart attitudes play an identical role in the individuals' mental economy. If individuals adopt a counterpart attitude toward counterpart propositions, I will say that they are in counterpart mental states.

Note that the easiest way of making sense of counterpart propositions in the way suggested by Conee and Madison would be by appeal to the notion of narrow content: if counterpart propositions are narrowly-individuated, the propositions embedded in Gary's and Barry's mental states would be "distinguished by nothing but the environment, if at all", and would be "as similar as possible while still being non-identical".

Relatedly, the most straightforward way of making sense of counterpart attitudes and counterpart mental states would be by reference to narrowly-individuated attitudes and states. If we conceive of factive attitudes as having a non-factive component, a factive attitude differs from its non-factive counterpart in an environmentally determined way, if at all, and will play the same role in the individuals' mental economy. The most

[39] While Audi proposes a hybrid access internalist account, he does not speak of counterpart propositions or states explicitly.

straightforward way of making sense of counterpart mental states would be by reference to purely internal states that, by hypothesis, may be shared by distinct broad states.

If the only way to spell out the counterpart relation is by reference to some narrowly-individuated property, however, the counterpart epistemic internalist proposal is in trouble. For even if many state externalists accept that some states may be narrow (such as pain states or other sensation states), many content externalists are sceptical that anything content-like is determined purely internally (see Adams et al. 1990; Yli-Vakkuri & Hawthorne 2018). Similarly, as discussed, many attitude externalists will reject compound views of factive attitudes (see Williamson 2000: Chap.1). So, if the only way to establish a counterpart relation between mental states is by reference to purely internal states, counterpart epistemic internalism is likely going to be rejected by state externalists. This is because, if counterpart mental states are defined by virtue of a purely internal state or property that counterparts share, justifying relations will exist only between purely internal mental states. Counterpart internalism would again be a concealed form of pure internalism.

2.3.1.3 Justification: internal justifying conditions

Madison proposes that Gary and Barry are justified in believing "exactly the same number of propositions and to exactly the same extent, on the basis of the same internally accessible grounds" (Madison 2009: 182). He continues:

> "what I am suggesting is that a specially indistinguishable experience, whether veridical or not, is shared by Oscar and Twin Oscar and is the basis of their justification for different propositions" (2009: 181).

Conee writes in a similar vein:

> "The claim of shared justification [between victims of extreme deceptions and their ordinary counterparts] is borne out by a variety of more specific internalist bases for justification. For instance, if the fundamental source of justification is experiential evidence that is

> shared by the counterpart pairs, then they have the same justification" (Conee 2007: 58).

On such a view, Oscar's internal state would justify him in believing that seawater is not drinking water but not in believing that twin seawater is not twin drinking water; the very same internal state would justify Twin Oscar in believing that twin seawater is not twin drinking water but not in believing that seawater is not drinking water.

What Madison and Conee must have in mind are purely internal contents, attitudes and experiences that, by hypothesis, are shared between Gary and Barry, and Oscar and Twin Oscar. The idea behind Madison's counterpart access internalism must be that Oscar and Twin Oscar have special access to the *same* facts about their evidence; and behind Conee's counterpart mentalism that Oscar and Twin Oscar are in the *same* relevant mental states.

While it may be that the same narrowly individuated states can justify counterpart propositions, this is not a view that interests us in the present context. According to Madison's and Conee's counterpart internalist views, an individual's external mental states do not partly determine whether the individual has justification for a certain belief. The justificatory work is entirely carried by the individual's purely internal states which, by hypothesis, are the same across distinct environments. Such a view is fake hybrid: while state externalism is accepted, it is sidelined from playing a role in the justification theory.

2.3.2 A better Counterpart Justification Thesis?

Here I will see whether the shortcomings of existing defences of the Counterpart Justification Thesis may be corrected. I will first have a look into the plausibility of the separability of individuation and justification and will find it only partly warranted (**§ 2.3.2.1**). Bracketing those mental states for which individuation and justification cannot be separated, I will see whether a proper counterpart account of justification of counterpart states can be given

that is consistent with state externalism, and will come to a negative conclusion (**§ 2.3.2.2**).

2.3.2.1 Limited separability of individuation and justification

The first assumption of counterpart epistemic internalism is that the enabling conditions of an individual's propositional attitudes may be partly determined by external factors while the justifying conditions are fully internally determined. The relevant dividing line between the internal and the external runs between environment-dependence and environment-independence: the environment-independence of justifying conditions is said to be consistent with the environment-dependence of enabling conditions.

An immediate issue is that if internality is understood as environment-independence in an unqualified way, the separability of individuation and justification is fairly obviously incoherent if hybrid epistemic internalism is assumed. Take hybrid mentalism. Mentalism identifies the justifying conditions with all of an individual's mental states. But, if state externalism is assumed, the individual's mental states are environment-dependent in an obvious way. As was briefly noted in §1.1.3.2 already, it seems clear that according to hybrid mentalism, "contingent and purely environmental variations" (Conee 2007: 59) will affect the justification of an individual's mental states by virtue of directly affecting the mental states themselves.

The same issue is likely to come up for hybrid access internalism. If some mental states are externally-individuated, it is likely that individuals have ready access to some of their external states. This is going to affect to which facts about their evidence individuals have ready access (e.g., Gary is likely to have ready access to the fact that he can see his hands whereas Barry will not have ready access to this fact). Again, the justifying conditions that hybrid access internalism specifies would seem to be environment-dependent in virtue of including environment-dependent mental states.

It may be in view of this tension that counterpart internalists like Madison and Conee restrict an individual's justifying conditions to the individual's purely internal mental states.

But the tension can be resolved if the sort of environment-dependence is qualified for individuation and justification, respectively. For note that a state can be environment-dependent in two distinct ways: a state can be environment-dependent in virtue of being *individuated* by the environment or in virtue of being *connected to the truth or to knowledge*. If the individuation of mental states is environment-dependent in one sense, and justification is environment-independent in a different sense, there may be no issue if the justifying conditions directly supervene on the enabling conditions. More precisely, counterpart epistemic internalists may suggest that the environment-dependence of a state in virtue of being *individuated* by the environment is harmless as long as the state is not necessarily *connected to the truth or to knowledge* in virtue of being externally-individuated. The incoherence would only arise out of an equivocation between these two kinds of environment-dependence.

Epistemic internalists who are moved by the Equal Justification Thesis believe that justification is environment-independent in the sense that justification is not necessarily connected to the truth or to knowledge, or put differently, that the evidential support relation is non-contingent. If whether someone has justification to believe something is directly connected to how likely it is that their beliefs are accurate, Gary's and Barry's beliefs could not be equally justified. As discussed in the first chapter (§1.3.1.2; §1.3.2), the New Evil Demon problem is invoked by epistemic internalists in order to argue that even if an individual's beliefs are reliably false, they may be justified, and even be as justified as the beliefs of someone who is reliably accurate. Only if the evidential support relation is a necessary relation, systematically deceived individuals could be as justified in their beliefs as their undeceived counterparts. So, epistemic internalists reject that there is a necessary truth-connection of justification.

For comparison, consider externalist justification theories. In externalist justification theories, justification is not independent of "contingent or purely environmental variations". Take reliabilism. Whether a belief is justified according to reliabilists depends on the environment because it depends on the environment whether a specific belief-forming process is

reliable.[40] Or consider the proper-functionalist account of epistemic justification defended by Bergmann (2006). According to the proper-functionalist account of justification, a belief can be justified only if the belief is the product of cognitive faculties that are functioning properly in an environment in which those faculties will reliably lead to the truth and for which that faculty was "designed" to function. Again, justification is not independent of "contingent or purely environmental variations" as it depends on the environment whether a specific faculty is functioning properly. Or consider the knowledge account of epistemic justification defended by Sutton (2005, 2007). According to the knowledge account of justification, a belief can be justified only if it constitutes knowledge. Again, justification is not independent of "contingent or purely environmental variations" as it depends on the environment whether you know that *p*. The same applies to Williamson's theory of justification that identifies your evidence with the content of everything you know (2000: Chap.9). The extent to which your belief is justified is the likelihood of the belief being true given everything you know. In virtue of being relative to your knowledge, whether you have justification to believe some *p* will again be dependent on the environment according to Williamson's account. It is signature trait of externalist justification theories that they make the justification of a belief dependent on its relation to processes or other beliefs that are connected in the right way to the contingent reality of how things are.

Silins proposes to understand the degree of truth-connection that is built into a justification theory as the defining feature that sets apart externalist from internalist justification theories: internalist justification theories are unified by their commitment to the absence of a necessary truth-connection of justification whereas different externalist justification theories endorse varying degrees of truth-connection of justification (Silins 2020: 6f.).

Any epistemic internalist, including hybrid epistemic internalists, who is motivated by the Equal Justification Thesis crucially cares about showing

[40] Nevertheless do forms of reliabilism other than simple reliabilism lead only to a weak truth-connection since even the beliefs that Barry forms in a really bad case may be reliably formed in virtue of the fact that the belief-forming processes are reliable in the good case (see Goldman 1986). Similar considerations apply to proper function externalism.

that justification is environment-independent in the sense that it is not necessarily connected to the truth. From the perspective of epistemic internalists, truth-connection is the harmful sort of environment-dependence because it characterises the fault line between epistemic internalism and externalism. When Conee writes that hybrid mentalism, while being "inclusive", is not non-committal insofar as it excludes justification from "contingent and purely environmental variations", his target is externalist justification theories, not state externalist theories. The sort of environment-dependence introduced by state externalism may be harmless, from the perspective of epistemic internalism, if the state is not necessarily *connected to the truth or to knowledge* in virtue of being externally-individuated. So if it is the case that an individual's evidence and mental states may be partly individuated by the environment without *ipso facto* being connected to the truth, state externalism does not introduce the sort of environment-dependence that would threaten to turn hybrid epistemic internalism into a non-internalist justification theory. In other words, when counterpart epistemic internalists argue that an individual's enabling conditions could be partly externally determined while the justifying conditions remain fully internally determined, what they mean is that whether an individual has justification for certain propositions is not necessarily connected to how likely it is that their beliefs are accurate by virtue of the fact that their mental states may be partly individuated by the environment.

Are they right? Is environment-dependence in terms of individuation consistent with environment-independence in terms of a truth-connection? It is evident that it is not in the case of factive states that combine the two forms of environment-dependence by being individuated *in virtue of* their truth-connection. That is presumably the deeper reason why the hybrid mentalists considered in §2.2.2 have sought to exclude factive mental states from their preferred way of setting up the supervenience base of (doxastic) justification. Singular thoughts pose a similar problem: as the thought is partly individuated in virtue of being related to a specific object, the individual could not be in the same mental state if the state was not at least partly accurate. According to certain singular externalists, this holds even for cases in which the individual

is wrong about some aspects of the object. Naïve realists, for example, group illusions together with veridical perceptions, and not with hallucinations, with respect to their individuation conditions: an individual's perceptual state is still partly individuated by the object if the individual has an experience as of a property of an object that the object in fact lacks (Martin 2002).

For non-factive states with contents that are not directly referential, the situation is more complex. It is clear that someone can have false beliefs while using externally-individuated natural kind concepts or socially-individuated concepts. Think of Barry in a recent bad case: Barry will have many false beliefs with externally-individuated contents, e.g. the belief that he just drank a glass of water. But it is not compatible with natural kind externalism or social kind externalism that an individual forms dramatically false beliefs for an extended period of time. In order to think about water given natural kind externalism, for example, one must have causally interacted with water at some point or someone in one's linguistic community must have interacted with water at some point. In order to think about arthritis, one must be partially right about how "arthritis" is defined in one's linguistic community. One could not, for example, believe that "arthritis" refers to a sort of fruit juice and successfully employ the concept ARTHRITIS. Nevertheless, it is compatible with natural kind externalism that one forms false beliefs and even forms reliably false beliefs for a limited period of time. It is compatible with social externalism that one forms false beliefs and even forms reliably false beliefs as long as one is not grossly mistaken about the meaning of the concepts employed. So it is correct that enabling and justifying conditions can be separated in the sense that states whose contents are externally-individuated according to natural kind externalism and social externalism are not *ipso facto* accurate, not even partially so.

But the separability of individuation and justification holds for fewer states and in more limited circumstances than counterpart epistemic internalists seem to think. The background assumption of counterpart epistemic internalists was that it would follow from the separability of justifying conditions and enabling conditions, that the same justifying conditions could occur across varying environments. Nothing that has been

said so far suggests that this is the case: even if the justifying conditions identified by hybrid epistemic internalism are partially environment-independent in the sense that they do not bear a necessary connection to the truth, they still directly supervene on conditions that are individuated by distinct environments. In order to establish the Counterpart Justification Thesis, it needs to be shown that a meaningful correspondence between justifying conditions across environments can be formulated. I will argue next that no such counterpart-relation could be formulated in a way that is consistent with state externalism.

2.3.2.2 Problems with the counterpart relation

Interestingly, proponents of counterpart access internalism do not consider the possibility of *counterpart justifying conditions* on which the justification of the counterpart states would supervene. But the initial characterisation of the the counterpart relation between Barry's and Gary's epistemic situations suggested that their counterpart beliefs would be justified by counterpart justifying conditions. Recall, when Gary has justification to believe that seawater is not drinking water because he drank it and it was too salty, then Barry has justification to believe that seavater is not drinking vater because he vat-drank seavater and it was too vat-salty. It seems natural, therefore, to propose a thoroughly counterpart-theoretic version of epistemic internalism where counterpart justifying conditions justify counterpart mental states across distinct environments.

We saw earlier that the challenge for spelling out the notion of counterpart mental states in a way that is consistent with state externalism is to do so without reference to some shared narrow content or shared purely internal state. What are the challenges for spelling out a notion of counterpart justifying conditions that is consistent with state externalism? Recall, for mentalism, the justifying conditions of an individual's beliefs is the totality of the individual's mental states. So, the challenge for giving an account of counterpart justifying conditions will also consist in giving a proper counterpart account of mental states. For access internalism, the justifying

conditions are facts about one's evidence to which one has ready access. If individuals have ready access to counterpart facts about their evidence, the justification of their counterpart mental states will supervene on counterpart justifying conditions. As was said in chapter 1 (§1.3.1.3), facts about one's evidence will be facts about one's mental states, such as facts about what one sees, knows, or justifiably believes. For example, Gary will have ready access to the fact that he knows that seawater tasted salty, whereas Barry will have ready access to the fact that he knows that seavater tasted vat-salty. So, the challenge for giving an account of counterpart justifying conditions in an hybrid access internalist account will again consist in giving a proper counterpart account of mental states. So whether one can spell out a notion of counterpart states as well as of counterpart justifying conditions that is consistent with state externalism will depend in both cases on whether the counterpart relation between mental states can be spelled out without reference to some shared purely internal state.

One way in which one may try to define the counterpart relation without reference to some shared purely internal state is by suggesting that occupying *counterpart functional roles*. Counterpart functional roles are constituted by the counterpart relation between the states and behaviours that tend to cause and be caused by the relevant states. If a counterpart state causes belief A (e.g., the belief that I should get some water), say, its counterpart state causes the counterpart belief A* (e.g., the belief that I should get some twin water); where one belief state causes action B (e.g., drinking water), its counterpart causes the counterpart action B* (e.g., drinking twin water). Regularly being caused by A and causing B would define a state's functional role F, and regularly being caused by A* and causing B* would define a state's functional role F*.

What defines the counterpart relation between F and F*? Here is one proposal. Distinct mental states occupy counterpart functional roles if and only if, were one mental state to occur in the counterpart environment, it would have the counterpart effects and it would be caused by the counterpart causes. For example, according to this proposal, Oscar's belief that he is thirsty for some water is a counterpart state to Twin Oscar's belief that he is thirsty for

some twin water, because if Oscar travelled to Twin Earth, and said that he is thirsty there, he would get a glass of twin water, and if Twin Oscar travelled to Earth and said that he is thirsty there, he would get a glass of water.

However, the fact that Oscar's belief would occupy the counterpart functional role to Twin Oscar's belief in this way if they swapped places, merely shows, according to internalists about functional roles, that Oscar's water-thoughts and Twin Oscar's twin water-thoughts occupy the *same* functional role in their respective mental economies. Internalists like Fodor, for example, argue that we should, and would generally in the natural sciences, assess causal powers across environments (see Fodor in Fodor and Martin 1986). It is precisely because Oscar gets a glass of twin water if he says that he is thirsty on Twin Earth, and *mutatis mutandis* for Twin Oscar, that Oscar's water-thoughts have the *same causal powers* as Twin Oscar's twin water-thoughts according to Fodor.

Externalists about functional roles will resist Fodor's claim. Burge and Davies, for example, object that it would be question-begging against the state externalist to suggest that the environment is dispensable when typing causal powers, and point out that in fact it is not considered dispensable in many special sciences (see Davies in Fodor and Davies 1986; Burge 1989). It makes little sense, for example, to abstract away from the fact that planets exist as part of a solar system in a description of a planet's causal powers (Davies 1986: 270). Similarly, they suggest that the state externalist proposal is precisely that the environment is indispensable for the individuation of causal powers in the special science of psychology. Externalists about functional roles will therefore insist that Oscar's water-thoughts and Twin Oscar's twin water-thoughts will exhibit distinct, externally-individuated functional roles (see also Block 1990).[41]

But by insisting that Oscar's and Twin Oscar's mental states occupy distinct functional roles, the externalist about functional roles has not told us in what sense the functional role that Oscar's beliefs about water occupy is a

[41] A closely related proposal defines the counterpart relation in behavourial terms whereby counterpart mental states are those states that cause counterpart types of behaviour. The behavourial definition attracts the same problems as the functional definition.

counterpart to the functional role that Twin Oscar's beliefs about twin water occupy, but is not a counterpart to the functional role that Twin Oscar's beliefs about juice occupy, for example. An externalist proposal that is consistent with the externalist claim that Oscar's and Twin Oscar's mental states occupy distinct functional roles is that the functional roles are *indiscriminable* in some sense to Oscar and Twin Oscar. If Oscar travelled to Twin Earth, for example, assuming that he is ignorant about his travels, and stayed there for a sufficient amount of time, he would not seem able to notice that he now has beliefs about twin water that cause behaviours related to twin water etc. The same considerations, *mutatis mutandis*, apply to Twin Oscar. So an externalist about functional roles may suggest that Oscar's water-thoughts occupy a counterpart functional role to Twin Oscar's twin water-thoughts in virtue of the fact that their mental states could swap functional roles in a way that Oscar and Twin Oscar would be unable to notice.

The most straightforward ways in which one may define the counterpart relation between mental states appeal to some narrowly-individuated property, such as a narrow content or a functional role that counterpart states would share. But many state externalists will reject that Gary's and Barry's mental states or Oscar's and Twin Oscar's mental states, respectively, have these properties in common. It is consistent with state externalists, however, that distinct mental states may be indiscriminable in some sense to the individuals involved. It will be worth inquiring whether hybrid epistemic internalists stand a better chance to defend something similar to the Equal Justification Thesis in terms of indiscriminability.

2.4 Vindicating the Indiscriminable Justification Thesis

It is consistent with an externalist view on a mental condition that distinct mental conditions are indiscriminable to an individual.[42] Recall disjunctivists about perceptual states, for example. Disjunctivists believe that veridical perceptions and hallucinations belong to fundamentally different kinds of mental phenomena. But they do not deny that an individual could experience a veridical perception and a perfectly matching hallucination successively and be unable to discriminate between them. They just deny that the veridical perception and its perfectly matching hallucination belong to the same kind of state in virtue of being indiscriminable if presented successively.

Similarly, one may think that the important correspondence relation between the mental states of individuals like Gary and Barry, on the one hand, and Oscar and Twin Oscar, on the other hand, is that their mental states are *indiscriminable* in a certain way. Likewise, their evidence and the relevant facts about their evidence may be indiscriminable to them in a certain way. If that is the case, one could propose the following weak variant of the Equal Justification Thesis. Define the neologism "indiscriminables" as referring to individuals whose mental states are indiscriminable to them in a certain way, and who furthermore have ready access to indiscriminable facts about their evidence. Like this, indiscriminables will satisfy the relevant correspondence relation according to mentalists and access internalists. If Gary and Barry have indiscriminable mental states in some way, they are indiscriminables in the mentalist sense. If they have ready access to indiscriminable facts about their evidence, they will be indiscriminables in the access internalist sense.

The "*Indiscriminable Justification Thesis*" states that indiscriminables have justification to believe *indiscriminable* propositions to the same extent. The Indiscriminable Justification Thesis would be entailed by a weakened variant of epistemic internalism in the following way:

---

[42] I use "mental condition" as synonymous with "mental phenomenon" which could be some thought content, an attitude, a mental state, a phenomenal state or any other mental phenomenon.

(i) Indiscriminables have justification to believe indiscriminable propositions to the same extent (Indiscriminable epistemic internalism).
(ii) Gary in the good case and Barry in the bad case (Oscar and Twin Oscar) are indiscriminables.
(iii) So, Gary and Barry (Oscar and Twin Oscar) have justification to believe indiscriminable propositions to the same extent (Indiscriminable Justification Thesis).

The proposal is that hybrid epistemic internalism may entail the following thesis even if it did not entail the original Equal Justification Thesis or the Counterpart Justification Thesis:

Indiscriminable Justification Thesis: Take the indiscriminables, G and B. G lives in the good case, B in the bad case. G and B have justification to believe *indiscriminable* propositions to the same extent.

In the rest of this section, I will do two things. I will clarify a number of notions central to understanding the Indiscriminability Justification Thesis, and motivate the idea that Gary and Barry, and Oscar and Twin Oscar, are indiscriminables, respectively (**§2.4.1**). However, the question whether Gary's and Barry's, or Oscar's and Twin Oscar's mental states are indiscriminable in a certain way is the central issue of the access problem. The next chapter 3 is dedicated in detail to the question whether Oscar is able to discriminate his actual water-thoughts from his counterfactual twin water-thoughts. In chapter 3, I will able to state precisely in what way Gary and Barry are indiscriminables.

For the rest of this section, I will assume that Gary and Barry are indiscriminables. According to the Indiscriminable Justification Thesis, they have justification to believe indiscriminable propositions to the same extent. I will motivate the Indiscriminable Justification Thesis and point out that the it yields a number of central epistemic internalist intuitions. But the

Indiscriminable Justification Thesis encounters the "*problem of inability*": in some cases, individuals are unable to readily discriminate between mental states and bodies of evidence, and yet they do not seem to have justification to believe indiscriminable propositions to the same extent, not even to the minds of hybrid epistemic internalists (**§2.4.2**). A very similar problem of inability is one of the central issues of chapter 4. At the end of chapter 4, I will able to state precisely in what way the Indiscriminable Justification Thesis has to be qualified in order to constitute a plausible solution to the equal justification problem.

2.4.1 The problem of presentations

To discriminate between two individuals is a sort of cognitive activity. More precisely, it is the cognitive activity of telling apart individual *a* from individual *b*, that is of activating knowledge that *a* is distinct from *b* (Williamson 1990: 7). Further, discrimination is relative to modes of presentation: *a* and *b* may be discriminable under some mode of presentation and indiscriminable under some other presentation. For example, while at a party, you may be able to readily discriminate the sandal-wearer from the man-bun wearer, but you may not be able to tell the writer from the yogi. What modes of presentation are available in a context will depend on the kind of evidence available to one in that context on which one can base one's judgment of distinctness.

What is the relevant mode of presentation when we are asking whether Gary and Barry (Oscar and Twin Oscar) are in indiscriminable mental states? For now, we have been supposing that Gary lives in the good case and Barry lives in the bad case. They are only ever in the mental states that they are in in their respective environments. Gary's and Barry's mental states do not pose an ordinary problem of discrimination: they are not like pairs of apples that they could hold in front of them in order to see whether they are able to discriminate between them. Call this the *problem of presentations*. The

problem of presentations is one of the problems that I will discuss in detail in the next chapter.

For now, let me adapt a classic variant of the original Twin Earth scenario to Gary and Barry to motivate the claim that Gary and Barry are indiscriminables (see Burge 1988; Boghossian 1989). Call the following a *switching case*.

Suppose that Gary leaves the good case and is transported to the bad case, i.e. the vat-verse. So Gary is envatted (and let us suppose that it is the whole Gary and not just his brain who gets envatted, for simplicity). The Earth and the vat-verse are distinguished by nothing but the fact that the fundamental reality in the vat-verse is made up of electronic signals. Gary starts to interact with vater, vat-washes with vater, vat-drinks it etc. Gary is not told about the fact that he has been envatted. As far as Gary is aware, he lives on Earth and uses the word "water" in agreement with his fellow inhabitants on Earth. But in fact, after some time, Gary expresses thoughts on vater when he uses the word "water". In the vat, Gary believes a proposition that he would express by "seawater is not drinking water". Gary thereby expresses a proposition about seavater and vater. But before the envatment, Gary believed the proposition that seawater is not drinking water that he would have expressed in the same way.

Is Gary able to discriminate his current vater-thought from his past water-thought that he would have expressed in the same way? It does not seem that he is. Vater and water only differ with respect to their fundamental structure. Nothing in Gary's experiential history gives Gary reasons to believe that he is presented with a different substance now than he was presented with in the past assuming that Gary is chemically ignorant. The same would seem to apply to his current vater-thoughts and his past water-thoughts. In order to readily know that he is currently thinking vater-thoughts, one may think that Gary must be able to readily discriminate the case in which his thought involves the concept VATER from the case in which it involves the concept WATER. But, by hypothesis, there is nothing in Gary's experiential history that provides him with reasons to readily discriminate the concept VATER from the concept WATER. Vater and water are perfect duplicates, except that

one is electronic and the other is an organic liquid. So, Gary would seem unable to discriminate his current vater-thoughts from his past water-thoughts when they are presented to him readily.

We could of course have told a similar story about Barry in which Barry would have been de-vatted and had continued to live on Earth. Essentially the same story has been told about Oscar and Twin Oscar in the so-called "slow-switching case" that I will briefly introduce in the next chapter (see Burge 1988; Boghossian 1989).

Note that, in a switching case, we are imagining that an individual moves between a good case and a bad case, or between their actual case and a twin case. In chapter 1, I showed that any state externalist position can be formulated with the help of a twin case (§1.2). So, we could describe a similar switching case for any state that is externally-individuated according to state externalists. In this chapter, I argued that there is a unity between twin cases and really bad cases (§2.2.1): mental states that will differ in and out of twin cases, will differ in between the good and really bad cases. Further, for those states that are not externally-individuated, i.e. for Gary's purely internal states, Barry will be in the same purely internal states in virtue of the fact that he is Gary's physical and phenomenal duplicate. So if Gary's mental states are readily indiscriminable from Barry's mental states in switching cases (and Oscar's from Twin Oscar's), as seems plausible, it is furthermore likely that *all* of their mental states are readily indiscriminable to them in switching cases, either in virtue of how they are presented to them or in virtue of the reflexivity of indiscriminability. It seems furthermore likely that they will have ready access to facts about their evidence that are not readily discriminable to them. For example, Gary, still in the good case, has ready access to the memory that seawater tasted extremely salty the last time he tried it. After the envatment, Gary may vat-taste seavater. He will have ready access to the memory that seavater vat-tasted extremely vat-salty the last time he vat-tried it. But nothing in Gary's experiential history provides him with reasons to readily discriminate between these two memories. It is plausible that Gary and Barry are indiscriminables in the mentalist as well as in the access internalist sense.

We will see in chapter 3 that a bit more argument is required in order to successfully show that Gary's and Barry's (Oscar's and Twin Oscar's) mental states are readily indiscriminable to them in switching cases given the appropriate modes of presentation. But I will argue there that the first impression sketched here is essentially correct.

2.4.2 The problem of inability

For the rest of this section, let us assume that Gary and Barry are indiscriminables. Is it plausible that Gary and Barry (or Gary after the envatment and sufficient time in the vat), and Oscar and Twin Oscar, respectively, have justification to believe indiscriminable propositions to the same extent in such circumstances?

Assume hybrid access internalism. Let us say that Gary in the good case has ready access to the fact that he can see his hands. On the basis of that fact, he has justification to believe that he has hands with a probability of 1. Gary, after the envatment and after sufficient time has passed, has ready access to the fact that he can vat-see his vat-hands. Plausibly, Gary in the vat will not be able to readily discriminate between the fact that he can vat-see his vat-hands from the fact that he can see his hands, if those facts are presented to him successively as "my current state of seeing my hands" and "my earlier state of seeing my hands". So, it seems plausible that Gary in the vat has justification to believe that he has vat-hands to the same extent, namely with a probability of 1, as Gary outside the vat has justification to believe that he has hands.

Of course, Gary does not have justification to believe that he has hands in the vat-verse, even though he may take himself to have justification to believe that (if he can still conceive of (organic) hands after the envatment, for discussion see §3.4.1). But he no longer believes that he has hands on the basis of his vat-ceptual state but he believes that he has vat-hands on the bases of the fact that he can vat-see his vat-hands.

Note that the Indiscriminable Justification Thesis would also seem to yield the correct result if we assume that individuals in the vat do not truly

believe propositions that are readily indiscriminable but distinct from those propositions that individuals outside the vat believe truly, but that they falsely believe the same propositions as individuals outside the vat. Suppose that after the envatment, Gary does not believe that he has vat-hands but falsely believes that he has hands, maybe because his natural kind-thoughts have not fully adapted to his new environment. Indiscriminability is reflexive: trivially, the proposition I HAVE HANDS is indiscriminable from itself. Further, Gary in the vat falsely believes that he hands on the basis of the fact that it seems to him that he can see his hands. But the fact that it presently seems to him that he can see his hands is readily indiscriminable to Gary from the earlier fact that he can see his hands if presented to him successively as "my current state of seeing my hands" and "my earlier state of seeing my hands". So even if Gary falsely believes that he has hands after the envatment, he has justification to believe the same proposition to the same extent inside and outside the vat according to the Indiscriminable Justification Thesis, which is the result important to epistemic internalists motivated by the Equal Justification Thesis.

Further, the Indiscriminable Justification Thesis safeguards intuitions that are dear to epistemic internalists. For example, a sceptic could rely on the Indiscriminable Justification Thesis to argue in the opposite direction: given that Gary after the long-time envatment has little to no justification to believe that he has vat-hands on the basis of facts about his evidence to which he has ready access, Gary, in the good case, who has ready access to facts about his evidence that are readily indiscriminable to him from the facts to which Gary in the vat has ready access, Gary has little to no justification to believe that he has hands. For better or worse, the Indiscriminable Justification Thesis still cuts both ways.

Let us consider a case that involves a sorites-series of mental states next. What we will see is that the Indiscriminable Justification Thesis requires an important qualification: it needs to control for cases in which an individual is unable to readily discriminate between epistemic situations (i.e. the relevant propositions and bodies of evidence) because they are making some sort of mistake. In these cases, not even epistemic internalists will want to say that the

individual has justification to believe indiscriminable propositions to the same extent.

Take an individual who is presented with coloured patches that incrementally change colour from purple to yellow over some period of time. Each situation in which the individual is looking at a coloured patch corresponds to a case. In the first case, the individual is looking at a purple patch and is in the perceptual state of seeing a purple patch. In the last case, the individual is looking at a yellow patch and is in the perceptual state of seeing a yellow patch. In between the first and the last case, the coloured patches change colour incrementally such that the colour of a patch in each case is readily indiscriminable to the individual from the colour of the immediately succeeding patch, if the cases are presented to them successively as "the shade a moment ago" and "THIS shade", where they refer to the colour of the patches in each case demonstratively.

Let us focus on two consecutive cases in the series, cases α and β in which the individual believes propositions α and β respectively. The proposition believed in α is readily indiscriminable to the individual from the proposition believed in β when presented to them successively. According to hybrid mentalists, the individual has justification to believe proposition α and proposition β to the same extent if their relevant mental states are readily indiscriminable to them in cases α and β. That is the case: the individual's perceptual state in α is readily indiscriminable to that individual from their perceptual state in β (given a successive mode of presentation). According to hybrid access internalists, the individual has justification to believe proposition α and proposition β to the same extent if the individual has ready access to readily indiscriminable facts about their evidence in cases α and β. Again that is the case: in case α, the individual has ready access to the fact that they can see that the coloured patch has a colour of THIS shade, and in case β, the individual has ready access to the fact that they can see that the coloured patch has the colour of THAT shade. Given that their perceptual states are readily indiscriminable to them in α and β, so will the facts be to which they have ready access. So, the individual will have justification to believe indiscriminable propositions to the same extent in the two consecutive cases α

and β according to hybrid access internalism. The Indiscriminable Justification Thesis yields those internalist results.

However, now consider the following variant of the case. It is known that individuals perform worse at virtually any task under extreme stress. Suppose someone is presented with the same sorites series of coloured patches and asked to press a button whenever they register a change in colour. Suppose further that they are simultaneously required to answer questions about an unrelated topic. It is plausible that the distraction may cause some of their perceptual states to be readily indiscriminable to them from a previous state when the same state would not have been readily indiscriminable to them from the earlier state had they been able to focus only on the change in colour.[43] In that case, the Indiscriminable Justification Thesis implausibly predicts that the individual has justification to believe indiscriminable propositions to the same extent in two cases when they are distracted and does not have justification to believe indiscriminable propositions to the same extent when they are not distracted. Not even epistemic internalists will believe that it is that easy to gain and lose justification for a proposition.

Note further that in the cases that motivate the Indiscriminable Justification Thesis, cases like Gary's and Barry's comparative epistemic situations, and Oscar's and Twin Oscar's comparative epistemic situations, it is not the case that their mental states and the facts about their evidence to which they have ready access are readily indiscriminable to them because Gary and Barry, or Oscar and Twin Oscar are making some sort of mistake, by being distracted, stressed or drunk or the like.

Some idealisation is required in order to identify the *right kind of inability* to readily discriminate between mental states and facts about one's evidence that would make it plausible that an individual has justification to believe indiscriminable propositions to the same extent in and out of twin

[43] There may be some ambiguity in the case with respect to whether the states are readily indiscriminable or whether the individual mistakenly judges them to be readily indiscriminable. But there will be cases where some temporary condition (like being drunk or tired) will affect the states one is in themselves and not just one's judgments about one's states.

cases and in and out of really bad cases. Call this the *problem of inability* for the Indiscriminable Justification Thesis.

A very similar problem of inability will be discussed at length in chapter 4 where I will defend a definition of the debate between internalists and externalists about mind in terms of an individual's ability for ready discrimination. At the end of chapter 4, I will be able to suggest a way in which the Indiscriminable Justification Thesis has to be qualified in order to constitute a plausible solution to the equal justification problem for hybrid epistemic internalism.

The Indiscriminable Justification Thesis is, of course, quite far off the initial motivation behind the Equal Justification Thesis, namely that individuals in really bad cases have justification to believe the very same things as individuals in the good case. This motivation is obviously not compatible with state externalism. But the Indiscriminable Justification Thesis yields a number of typical epistemic internalist intuitions. Importantly, individuals in bad cases are no less justified in believing what they believe in their epistemic situation than what their duplicates in the good case have justification to believe in their epistemic situation.

2.5 Conclusion

In this chapter, I looked into three ways that hybrid epistemic internalists may defend the claim that hybrid epistemic internalism entails the Equal Justification Thesis, or a variant.

The first strategy may be labelled as "faking it". Those accounts argued that hybrid epistemic internalism entails the original Equal Justification Thesis by paying lip service to state externalism while keeping it from coming into effect by stipulations and exclusions.

The second strategy may be labelled as "finding it". Those accounts tried to show that hybrid epistemic internalism entails the Counterpart Justification Thesis by virtue of a substantial counterpart-relation that would

exist between the mental states and the bodies of evidence of individuals in and out of really bad cases and in and out of twin cases. The issue was that many state externalists would reject that such a substantial counterpart-relation exists.

The last strategy may be labelled as "feigning it". These accounts argued that hybrid epistemic internalism entails the Indiscriminable Justification Thesis by virtue of the purely, epistemic relationship that distinct mental states and distinct bodies of evidence of individuals in really bad cases and in twin cases are readily indiscriminable to them from the mental states and bodies of evidence of their duplicates in the good case or in the actual environment. The Indiscriminable Justification Thesis is the weakest but most promising way in which hybrid epistemic internalists may motivate their view by a thesis reminding of the Equal Justification Thesis.

The Indiscriminable Justification Thesis works if duplicates in and out of really bad cases and in and out of twin cases are indiscriminables in a way that excludes certain mistakes from interfering. In the very last section of this thesis, I will be able to suggest a version of the Indiscriminable Justification Thesis that solves the equal justification problem for hybrid epistemic internalists. This is because, as we will see in the next chapter, solving the problem of presentations involves showing that state externalism leads to an access problem. And solving the problem of inability will be part of showing that it does so by definition. Both these solutions are necessary in order to formulate a properly qualified version of the equal justification thesis. What this means is that the equal justification problem can be solved in virtue of the fact that state externalism leads to an access problem, and that by definition.

Chapter 3

3 - The Discrimination Argument Against Ready Access

3.1 Introduction

State externalists have been worried about the compatibility of state externalism and access to one's external mental states in virtue of the so-called discrimination argument. Roughly, the discrimination argument claims that if state externalism is correct, individuals will sometimes fail to know that they are in a particular external state in virtue of the fact that they are unable to discriminate that state from a relevant alternative external state. There is the widely discussed discrimination argument against *special* access, but in this chapter I will focus on the discrimination argument I believe to be more pertinent, namely the discrimination argument against *ready* access to one's external mental states.

In response to the discrimination argument against ready access, there are two compatibilist strategies open to state externalists ('compatibilists' because they believe that state externalism and ready access to one's occurrent external states are compatible). Proponents of the '*Even-if-strategy'* argue that individuals have ready access to their occurrent external states even if they are unable to discriminate a particular occurrent external state from a relevant alternative external state. Proponents of the '*Don't-worry-strategy*' reject the cogency of the discrimination argument itself. Here I will respond to the *Don't-worry-strategy*.

There are two reasons for which state externalists may reject the discrimination argument. They may question whether there are appropriate modes of presentation under which occurrent external states are indiscriminable from counterfactual external states. Call this the *objection from presentations*. Others may argue that the circumstances in which indiscriminable alternative external states are relevant are not normally relevant to us. Call this the *objection from relevance*.

In this chapter, I will defend the discrimination argument against both objections. The discrimination argument is sound, and the circumstances in which the discrimination argument is pertinent are furthermore sometimes relevant to us. The result is that, if state externalism is correct, we sometimes do not know that we are presently in some particular external state rather than

some relevant alternative state unless we change our epistemic situation significantly.

The fact that state externalism limits the ready access individuals have to their mental states in this way is likely to be a problem for hybrid access internalism, that combines state externalism with a view on epistemic justification according to which whether an individual has justification to believe some *p* supervenes on those facts to which the individual has ready access. In order to show that state externalism is compatible with ready access to one's mental states in certain circumstances, compatibilists fully depend on the success of the *Even-if-strategy*.

The plan of the chapter is as follows. I will introduce the discrimination argument in **§3.2**. I'll comment on the two compatibilist strategies to the discrimination argument in **§3.3**. In **§3.4**, I will discuss the objection from presentations and show that there are two modes of presentation under which present external states are indiscriminable from relevant counterfactual mental states in certain circumstances. In **§3.5**, I will respond to the objection from relevance by showing that these circumstances are sometimes relevant to us. **§3.6** concludes.

3.2 The Discrimination Argument

I will lay out in more detail how state externalism would seem to undermine an individual's ready access to their external mental states in certain circumstances (**§3.2.1**). I'll contrast the discrimination against *ready* access with the discrimination argument against *special* access. We will see that the discrimination argument against ready access identifies the central assumption behind the discrimination argument in a way the discrimination argument against special access does not (**§3.2.2**).

3.2.1 The discrimination argument against ready access

Think of Oscar in the classic Twin Earth scenario. One may think that in order for Oscar to know that he is thinking a water-thought, he would have to know that his thought involves the concept WATER and not the concept TWIN WATER. But in order to know that his thought involves the concept WATER and not the concept TWIN WATER, Oscar will have to research the environment. This is because Oscar would have to know that his thoughts are typically *de re* H_2O and not typically *de re* XYZ, and he would not be able to know that without first investigating his environment. In that case, Oscar would not be able to readily know that he is thinking a water-thought.

As has been noted, this line of reasoning is just too quick (see Boghossian 1989: 158). In the classic Twin Earth case, Oscar never leaves Earth and Twin Oscar never leaves Twin Earth. There is no risk that Oscar would be teleported to Twin Earth and come across twin water. According to ordinary standards of knowledge, Oscar does not have to be able to discriminate his water-thoughts from thoughts about twin water he is at no risk of having in order to know that he is thinking a water-thought.[44]

But we already encountered switching cases in the last chapter. In a switching case, an individual is "switched" between two different environments such as a good case and a really bad case, or their original environment and a twin environment. In the classic switching case version from the Twin Earth scenario, Oscar is slowly switched between Earth and Twin Earth (Burge 1988: 652; Boghossian 1989: 158). We suppose that Twin Earth is an actual planet in another part of our galaxy and that, without his knowledge, Oscar has for some time been undergoing a series of switches between Earth and Twin Earth. Earth and Twin Earth are, as usual, distinguished by nothing but the molecular structure of their dominant liquid of which Oscar, as well as the population on each planet, is ignorant. At each stop, Oscar stays long enough to acquire the concept that the local population would express with "water" by successfully interacting with the dominant

[44] Not according to sceptical standards for knowledge, of course, as the sceptic requires that individuals are able to discriminate the current case from any alternative case in order to know things in the actual case.

liquid for a sufficient period of time. Currently on Twin Earth, Oscar believes a proposition that he would express by "seawater is not drinking water". In such switching case, Oscar would not seem to be able to readily know that he is currently thinking a twin water-thought because he is unable to discriminate the case in which his thought involves the concept TWIN WATER from the relevant alternative case in which his thought involves the concept WATER before he finds out more about his environment.

As pointed out in the previous chapter, switching cases can be formulated for any state that is externally-individuated according to state externalists in virtue of the fact that switching cases are twin cases become relevant. We can define switching cases as follows:

> *Switching case:*
> *Define a case as as a total state of a possible system consisting of a subject *S* in a mental state *M* paired with an environment *E* at a time *t*.
> *Take a case α and the *relevant* case β: S_α is in M_α in α and S_β is in M_β in β.
> *Case α is a *switching case* to case β if and only if S_α and S_β are physically or phenomenally identical (or both).

If state externalism is true, moving from E_α to E_β will result in moving from M_α to M_β after a sufficient amount of time, and $M_\alpha \neq M_\beta$. I will also speak of "mental state switching": assuming state externalism, M_α switches to M_β if and only if the mental state change is caused by a change to the environmental feature identified as relevant by some state externalist view. Switching may occur slowly as in the case of natural kind thoughts, i.e. thoughts that embed a natural kind concept like WATER, because individuals will have to spend a substantial amount of time in the new environment before their natural kind concepts adapt to it. In other cases, switching will occur instantaneously, for

example if the object of a demonstrative thought switches, assuming singular externalism.[45]

The fully general discrimination argument against ready access to one's external mental states can be stated as follows:

(1) To readily know that S is in external state M_α, S must be able to readily discriminate the case α in which S is in M_α from any relevant alternative case in which S is not in M_α.
(2) In a switching case, the case β in which S is in the external state M_β is a relevant alternative to the case α.
(3) In a switching case, S is unable to readily discriminate the case α from the relevant alternative case β.
(C) So, S does not readily know that S is in M_α in a switching case.

Most theorists accept that it would be a serious objection to state externalism if it was a consequence of state externalism that individuals would have to research the environment in order to know that they are in some occurrent mental state (e.g., Davidson 1987; Burge 1988; Heil 1988; Falvey and Owens 1994; Gibbons 1996; McLaughlin and Tye 1998).

The two problematic premises of the discrimination argument are premises (1) and (3), as, by hypothesis, the alternative mental state is relevant in a switching case.

Premise (1) is the discrimination condition on knowledge applied to ready knowledge of one's mental states. As we will see in **§3.3**, proponents of the *Even-if-strategy* question the assumption that the discrimination condition on knowledge applies to ready knowledge of one's mental states. Whether a compatibilist about state externalism and ready access opts for the *Even-if-strategy* or for the *Don't-worry-strategy* will depend on how likely they think it is that the discrimination condition on knowledge applies to self-knowledge.

[45] For this reason, I chose the inclusive label of "switching cases" over the more common label of "slow-switching cases": as we will see in detail in §3.5, not all switching cases are slow.

One way to resist premise (3) is to argue that, in a switching case, there is not an appropriate, i.e. theoretically significant, mode of presentation under which an individual's occurrent external states are not readily discriminable to them from some relevant alternative states. I call this the "objection from presentations" to the discrimination argument. In **§3.4**, I will show that there are two appropriate modes of presentation under which external mental states are not readily discriminable to an individual in a switching case.

I'll discuss the plausibility of premises (1) and (3) with a focus on the classic switching case. If (1)-(3) can be established, it follows that anyone who, like Oscar, is switched between Earth and Twin Earth does not readily know that they are in a water-state rather than a twin water-state. So what? None of us is being switched between Earth and Twin Earth. Another way a proponent of the *Don't-worry-strategy* can resist the discrimination argument is by arguing that switching cases are not relevant to *us*. In **§3.5**, I'll argue that switching cases are sometimes relevant to us. **§3.6** concludes.

3.2.2 Contrast: the discrimination argument against special access

The discrimination argument against *ready* access to external mental states is similar to the better known discrimination argument against *special access* (for discussion see, e.g. Burge 1988; Boghossian 1989; Heil 1988; Brueckner 1990; Nuccetelli 1993; Falvey and Owens 1994; Brown 1995; Ludlow 1995; Gibbons 1996; McLaughlin and Tye 1998; Farkas 2008; Goldberg 2015). Let us abbreviate the former as "the discrimination argument$_R$" and the latter as the "discrimination argument$_S$" for the duration of this section. The arguments, although similar, differ in crucial respects. It will be helpful to briefly point out where.

As discussed in chapter 1, we have special access to some fact that p if and only if we know that p in a peculiar and privileged way. To have special access to some fact is one way to have ready access to that fact. This is because ready access is permissive of many ways of knowing things: as long as a fact is accessible without significantly changing one's epistemic situation,

a number of methods, empirical as well as special, can contribute to making that fact readily accessible. Special access, in contrast, has often been characterised in a way that suggests a contrast to empirical forms of access, e.g. as "non-empirical", "non-observational", "a priori", or "reflective" (Chisholm 1977; Bernecker and Dretske 2000; Pryor 2001; Huemer 2007). In other words, special access entails ready access but not the other way around. So, if the discrimination argument$_R$ is sound, then so is the discrimination argument$_S$.

If the discrimination argument$_R$ is sound, it shows that the discrimination argument$_S$ emphasises the wrong fault line.

The discrimination argument$_S$ takes the access problem for state externalism to consist in the fact that it would make knowledge of some of our mental states depend on empirical evidence. The access problem would consist in the fact that we assume to have non-empirical, non-observational or *a priori* knowledge of our mental states, but state externalism makes knowledge of some of our mental states depend on researching our environment.

By contrast, the discrimination argument$_R$ sees the source of the access problem as the fact that state externalism makes knowledge of some of our mental states depend on facts that we are not in a position to know in a given context without changing our epistemic situation significantly. The fact that our mental states may vary with factors that we are not in a position to know readily seems to be a consequence of the fact that state externalism considers some of our mental states to supervene on things such as natural kinds or expert definitions in our linguistic community, which, in some circumstances, we will be able to know only by getting further evidence about our environment or our linguistic community. Here the access problem consists in the fact that state externalism makes knowledge of some of our mental states depend on *more* empirical evidence than we have in a particular context. So, according to the discrimination argument$_R$, the access problem does not consist in the fact that empirical evidence may be involved at all in how we know that we are in a particular mental state, only in the fact that, in some

circumstances, we would need empirical evidence *beyond the readily accessible evidence*.

An advantage of formulating the discrimination argument while making the appeal to special access dispensable is that the argument will speak to those state externalists who de-emphasise the role of special access in their philosophy of mind and justification theories (Williamson 2000; Gibbons 2006; Das and Salow 2016). Many state externalists are attracted to deflationary theories of self-knowledge that consider knowledge of one's surroundings prior and knowledge of one's mind derivative on the knowledge of one's surroundings (Evans 1982; Das and Salow 2016; Byrne 2018). But, however deflationary one's theory of self-knowledge is, one faces an access problem if state externalism is correct. This is because any self-knowledge theorist assumes that the kind of mental states that are being discussed in the discrimination argument, mental states such as Oscar's occurrent thought that seawater is not drinking water, are among the facts that are readily accessible to an individual. The discrimination argument$_R$, if it is correct, shows that it is solely in virtue of state externalism itself that an access problem ensues. No further assumptions about the specific mechanisms of self-knowledge are required.[46]

3.3 Two Compatibilist Strategies

The first premise of the discrimination argument against ready access is an application of the discrimination condition on knowledge to the case of ready knowledge of one's mental states (ready self-knowledge). I'll briefly motivate the discrimination condition in **§3.3.1**. Whether the discrimination condition on knowledge is plausible for empirical knowledge or not, proponents of the

[46] Note, though, that I am suggesting that special access is irrelevant only with respect to the so-called *achievement problem* of state externalism, which is the problem of how self-knowledge can be achieved given state externalism. For the so-called *consequence problem*, it is relevant that we often know that we are in a particular mental state specially: the consequence problem is the issue that we may have special access to our environment if we have special access to our external mental states (see McKinsey 1991).

compatibilist *Even-if strategy* believe that the case of self-knowledge is special insofar as individuals are in a position to know that they are in a particular mental state *even if* they are unable to discriminate the occurrent mental state from relevant alternative mental states (**§3.3.2)**. The disagreement between the *Even-if* and *Don't-worry* strategies is therefore also a disagreement about the right account of self-knowledge. Instead of engaging in that debate, I demonstrate in **§3.3.3** that the *Don't-worry-strategy* may be the safer bet for compatibilists, as the *Even-if-strategy* raises a number of serious issues.

3.3.1 The discrimination condition on knowledge

Recall the first premise of the discrimination argument:

> (1) To readily know that S is in M_α, S must be able to readily discriminate the case α in which S is in M_α from any relevant alternative case in which S is not in M_α.

The thought that motivates the first premise is the widespread idea that ordinary knowledge requires a certain kind of insulation or protection against the possibility of error. Some epistemologists have interpreted the requirement for protection from error as a requirement for certain discriminative abilities. The idea is that knowledge is closely connected with the notion of being able to discriminate when something is the case from when it is not the case (see, e.g., Goldman 1976, 1986; Dretske 1981; McGinn 1984; Lewis 1996).

Here is an example of the kind of intuitions that have led a number of philosophers to believe that the notion of knowledge is linked to certain discriminative abilities (for the example, see Brown 2004: 41). Compare a lay bird-watcher and an expert who are bird-watching in good light. Both of them form the correct perceptual belief that the bird in front of them is a crow. However, we would not say that the layperson knows that the bird in front of them is a crow if they cannot distinguish a crow from another bird abundant in the area, a rook for example. This is because, unlike the expert, in a

counterfactual case in which a rook is sitting in front of them, the layperson would falsely believe that the bird in front of them is a crow owing to their inability to discriminate a rook from a crow.

The notion of knowledge, however, does not seem to be linked to just any discriminative abilities. If there exists a rare bird very similar to a crow in a remote area in which the expert does not live and is unlikely to visit, we would say that they know that the bird is a crow even if they cannot discriminate the crow from the rare bird. This is because the expert is able to discriminate the case in which they are facing a crow from all *relevant* cases in which they are not facing a crow.

On the basis of such intuitions, a number of epistemologists endorse a discrimination condition on knowledge along the following lines:

> *Discrimination condition on knowledge*: S knows that *p* if and only if they can discriminate the actual case in which *p* is true from all relevant counterfactual cases in which *p* is false (Goldman 1986: 46).

3.3.2 The Even-if-strategy

While a number of influential epistemologists have interpreted the idea that knowledge requires a certain kind of protection from error as supporting a connection with certain discriminative abilities, other epistemologists have stressed a connection between knowledge and the reliability of one's belief-forming mechanism, or between knowledge and certain modal conditions to the same end.[47] A popular modal condition on knowledge is *safety* (see, e.g. Sosa 1999; Williamson 2000; Pritchard, 2005). Suppose one believes that *p* via method *m* at time *t*. We can then define:

---

[47] Traditional compatibilist accounts that were developed in the 80s and 90s are formulated in terms of reliability: reliably formed beliefs are protected from error in most cases. More recent compatibilist accounts like Das and Salow's view however appeal to safety. I focus on the safety-condition because it is clearer and more widely defended today.

> *Safety condition on knowledge*: An individual's belief that *p* formed via method *m* is safe at t if and only if there is no close possible world in which one believes that *p* via *m* at t and *p* is false.

The discrimination condition and the safety condition will generate the same verdicts on whether an individual knows a given *p* in most cases, but for slightly different reasons. What matters on the discrimination condition is whether an individual is *able to discriminate* the close counterfactual case in which the believed proposition is false from a case where it is not. The lay birdwatcher fails to know that the bird is a crow because they cannot discriminate a case in which the bird is a crow from a case in which it is a rook. What matters for the safety condition is whether an individual *falsely believes that p* in a close counterfactual scenario. The lay birdwatcher's belief that the bird is a crow is also unsafe because in a close counterfactual case it is a rook in front of them but they believe that it is a crow. If, however, S does not falsely believe that *p* in any of the close counterfactual cases, but truly believes a different proposition *p**, it is irrelevant on the safety condition that S is unable to discriminate those close counterfactual cases from the actual case even if *p* is false (but not believed) in some of those close counterfactual cases.

This difference between the discrimination condition and the safety condition turns out to be crucial if we consider whether Oscar knows that he is in a particular external mental state in the classic switching case. Recall: after having been switched, Oscar on Twin Earth believes that he believes that twin seawater is not twin drinking water, a belief that he would express as "I believe that seawater is not drinking water". Does Oscar know that he is presently thinking that twin seawater is not twin drinking water? The discrimination condition on knowledge says that Oscar does not know that he is thinking a twin water-thought if he is unable to discriminate his present twin water-thought from the relevant counterfactual water-thought that seawater is not drinking water. The safety condition on knowledge says that Oscar does not know that he is presently thinking a twin water-thought if he falsely believes that he is thinking a twin water-thought in a close counterfactual case.

Does Oscar falsely believe that he is thinking a twin water-thought in a relevant counterfactual case in a switching case? Oscar would falsely believe that he was thinking a twin water-thought if he believed that he was thinking a twin water-thought when in fact he was on Earth thinking a water-thought. But if Oscar were on Earth, thinking with Earth concepts, he would not presently have beliefs about twin water but about water. More precisely, if Oscar were on Earth, he would presently think that seawater was not drinking water. Furthermore, he would believe that he was presently thinking that seawater was not drinking water. But then Oscar does not falsely believe that he is thinking a twin water-thought in a relevant counterfactual case, but instead truly believes that he is thinking a water-thought. As mentioned, a true counterfactual belief in *p** does not undermine the safety of the actual belief that *p*.

Proponents of the *Even-if-strategy* reject the discrimination condition in the case of self-knowledge in favour of some other condition, such as the safety condition (see, e.g., Burge 1988; Heil 1988; Falvey and Owens 1994; Gibbons 1996; Jacobsen 1997; McLaughlin and Tye 1998; Das and Salow 2016). Their central claim is that individuals have access to their occurrent external states *even if* they are unable to discriminate a particular occurrent external state from a relevant alternative external state.

We may also say that they offer a *compatibilist* account of self-knowledge: state externalism is compatible with ready access to one's occurrent external states in virtue of the fact that, even in switching cases, an individual would not have a false second-order belief in a relevant counterfactual case but a distinct true second-order belief. I will sometimes speak of *discriminatory* self-knowledge, as opposed to *compatibilist* self-knowledge, that individuals lack when they are unable to readily discriminate between some occurrent state and a relevant alternative state. Compatibilist self-knowledge is compatible with the lack of discriminatory self-knowledge. However, I will mostly try to avoid these labels in order not to suggest that these are two different yet compatible kinds of self-knowledge. The respective proponents take themselves to offer the only correct account of self-knowledge.

Burge calls the judgments by which Oscar self-ascribes his presently entertained thoughts in the classic switching case "cogito-judgments" (1988: 658). In virtue of the fact that an individual cannot self-ascribe a present-tense thought content that they are not presently thinking, cogito-judgments are contextually self-verifying and thereby trivially safe. Importantly, the safety of cogito-judgments is compatible with the inability of the individual to readily discriminate the thought self-ascribed from distinct thoughts they would be thinking in a counterfactual case. Burge writes, for example "[The slow switch subject] would have no signs of the differences in his thoughts, no difference in the way things feel. . . . the person would be unable to discriminate between different mental events under the stated switching conditions" (1988: 653).[48] Even if Oscar has been switched and is unaware of the switch, he is able to think occurrently entertained thoughts self-ascriptively. He will, furthermore, be able to do so readily for, given certain background conditions such as possessing the required conceptual repertoire, being awake and not heavily intoxicated, adult humans are readily capable of self-ascribing thought contents. So, cogito-judgments would seem to provide ready self-knowledge of contents even in switching cases.

I presented the structure of compatibilist accounts of ready knowledge of externally-individuated thought contents. More recently, a structurally similar account has been defended for ready knowledge of factive states. Das and Salow (2016) argue that "we can base our beliefs about mental states on those mental states themselves (…) there is thus no risk of my going wrong, and nothing substantive is required to make my belief safe" (2016: 15). Roughly, Das and Salow argue as follows. According to them, we can infer that we are in a particular mental state by virtue of so-called transparent inferences. We may apply such a transparent inference directly to a particular factive state and transparently infer that we currently are in some factive state, for example, the state that we know that seawater is not drinking water. Since the inference is based directly on that factive state, Das and Salow argue that we would not draw the *same* kind of inference had we falsely inferred from a non-factive

[48] For similar comments, see Boghossian 1989: 160; Burge 1996: 95.

state that we are in a factive state. For example, according to them, had we inferred from the mere true belief that seawater is not drinking water that we know that seawater is not drinking water, we would not have drawn the *same* kind of inference. But, importantly, the kind of inference we draw fixes which possible cases are close cases: the close cases are selected only from the set of cases in which the belief is formed via the same method. So, Das and Salow argue, only cases where the second-order belief is also based on a factive state will be close cases. But if only cases in which the second-order belief is based on a factive state will be close, then no case in which we falsely believe that we are in that factive state could be close. So, our belief that we are in that factive state is safe. That is the rough structure of Das and Salow's compatibilist account of self-knowledge of factive attitudes.

The structural commonality with the compatibilist accounts of thought content is that self-knowledge of factive attitudes is argued to be trivially achieved, although in this case not in virtue of being contextually self-verifying, but in virtue of the specific structure of transparent inferences. Again, owing to the supposed fact that the self-knowledge is trivial, it is argued to be available in switching cases: as long as several background conditions about the individual hold (such as the capacity for self-reflection and inferential reasoning), adult humans are capable of transparently inferring that they are in a particular occurrent factive state even if they have been switched unbeknownst to them. The further ingenuity of Das and Salow's account, if correct, is that, in virtue of how methods are individuated and close cases selected by reference to method, any case in which the individual merely tries to transparently infer from a non-factive state that they are in a factive state is not a close case. But not even the discrimination condition requires that individuals have to be able to discriminate the actual case from irrelevant counterfactual cases in which they falsely believe the same proposition.

3.3.3 The Don't-worry-strategy

Some compatibilists do not find the *Even-if-strategy* against the discrimination argument very compelling. This is because one can raise a number of questions about compatibilist accounts of self-knowledge.

Consider, first, that the knowledge of thought content that is trivially safe in the way described by Burge accounts for very little of our self-knowledge. Consider Oscar, who, on Twin Earth, self-ascriptively judges "(with that very thought) I'm currently thinking that seawater is not drinking water" thereby coming to know that he is presently thinking a twin water-thought. Compare if Oscar had judged "I am annoyed that seawater is not drinking water". The fact that Oscar judges that he is annoyed that twin seawater is not twin drinking water does not make it true that Oscar is actually annoyed about it. Burge's account of self-knowledge does not provide ready access to the attitude one takes toward the thought content. This holds for standing as well as occurrent attitudes: if Oscar judges that he fears that twin seawater is not twin drinking water, it does not thereby become true that he has been fearing or that he is currently fearing it. The same applies if the slightest time difference comes in between the first-order thought and the self-ascribing judgment. Oscar's judgment that he just now thought that twin seawater is not twin drinking water is not self-verifying: it need not be true that Oscar had that thought a moment ago for him to judge now that he had it a moment ago. In other words, self-knowledge of thought contents is self-verifying only in the very special case where the second-order thought literally embeds the first-order thought about which it is thought.[49] The self-knowledge of thought content that is trivially achievable and available even in switching circumstances is limited to the readily available knowledge that one is thinking the content one is thinking. One may reasonably doubt that state externalists worried by the discrimination argument against ready access are going to be appeased by learning that state externalism does not jeopardise this sort of knowledge.

---

[49] See Boghossian (1989: 168ff.) for raising these issues.

As pointed out, the self-knowledge of factive attitudes, in the way Das and Salow defend it, is not contextually self-verifying: it certainly is not contradictory to suppose that a case in which one attempts to transparently infer that one is in a factive state from a non-factive state is close to a case in which one successfully infers that one is in a factive state from the state itself. Das and Salow's compatibilist account builds on a substantial account of how methods are individuated and close cases selected that may be false. So, while it is more likely that Das and Salow give an account of self-knowledge of attitudes, it is less likely that it succeeds in showing that self-knowledge is always achievable, let alone trivially so.

On the basis of these considerations, a number of self-knowledge theorists have doubted that compatibilist knowledge constitutes genuine self-knowledge (see, e.g., Boghossian 1989; Sawyer 1999; Goldberg 2000). For consider that Oscar's self-ascribed thought is true whichever thought, a twin water- or a water- second-order thought, is made true by his thinking. Oscar's cogito-judgements share that property with other contextually self-verifying judgments such as "I exist", "I am here" or "a thought was thought" that, whenever thought or uttered, are *ipso facto* true. But self-knowledge, such as the knowledge that you are annoyed that seawater is not drinking water, is not usually contextually self-verifying in this way.

The structural reason for these issues with compatibilist accounts of self-knowledge is that it is unclear to what extent the safety condition works as a requirement on knowledge independently of the discrimination condition. First, note that the satisfaction of the discrimination condition and of the safety condition are often interrelated by explanatory ties. It is often *in virtue of* lacking certain discriminative abilities that an individual's belief is not safe. The lay birdwatcher's belief that it is a crow in front of them is unsafe because it satisfies the following counterfactual:

> (C) If a rook had been in front of the lay birdwatcher, they would have falsely believed that it was a crow.

And it is often in virtue of having certain discriminative abilities that an individual's belief is safe. The expert birdwatcher's belief is safe because it satisfies the following counterfactual:

> (C*) If a rook had been in front of the expert birdwatcher, they would not have believed that it was a crow.

Second, the lack of discriminative abilities makes for an interesting difference in ways individuals satisfy the relevant counterfactuals. When we assess whether Oscar knows that he is thinking a particular thought, we are asking which of the following counterfactuals is correct:

> (C**) If Oscar had thought about twin water, he would have falsely believed that he was thinking about water.

> (C***) If Oscar had thought about twin water, he would not have believed that he was thinking about water.

Compatibilists argue that (C***) is the correct counterfactual. But notice the difference between (C*) and (C***). (C*) is true because the expert birdwatcher possesses the ability to discriminate between rooks and crows; in the expert's counterfactual scenario, a cognitive act of discrimination would explain the judgment that it is a rook and not a crow in front of them. But Oscar, by hypothesis, does not possess any experiential evidence that would indicate to him that he is thinking a water- rather than a twin water-thought, so if he is able to discriminate between them, this must be due to something else than experiential evidence. But according to most proponents of the *Even-if-strategy*, Oscar will satisfy (C***) *even if* he is unable to discriminate between water- and twin water-thoughts, and so without being able to perform any cognitive act of discrimination that would explain the truth of the counterfactual. Instead, Oscar would trivially satisfy (C***).

The classic switching case is one of the few cases where safety and discriminative abilities come apart.[50] To some, the difference between how the relevant individuals satisfy (C*) and (C***) indicates that something may be off with a purely safety-based account of self-knowledge. Systematic research into the connection between safety and discriminative abilities is yet to be forthcoming, although the beginnings are discernible (see, e.g., Goldberg 1999, 2006; Sawyer 2014, 2015; Callahan 2020).

The *Even-if-strategy* may not work because it could turn out that the discrimination condition applies to self-knowledge. As long as the separability of discriminative abilities and safety is unclear, it may be safer for compatibilists to try a more direct strategy against the discrimination argument. Proponents of the *Don't-worry-strategy* argue that state externalists do not have to worry about the discrimination argument, either because it is not sound or because it is sound but not relevant to us (see, e.g., Warfield 1997; Sawyer 1999; Brown 2004). I will defend the discrimination argument against both objections in the next two sections.

3.4 The Objection from Presentations

Discrimination is relative to modes of presentation: two individuals may be indiscriminable under some modes and discriminable under others. Before we know the modes of presentations under which an individual's external states are presented to them in a switching case, we cannot really tell whether they are unable to readily discriminate between them. What modes of presentation are available in switching cases depends on the further question of what exactly happens to the relevant mental states in a switching case. The brief introduction of switching cases in §3.2.1 glossed over the fact that theorists by no means agree on the conceptual effects of switching. As we will see, only

[50] Brown works out in detail how those compatibilist accounts that are aware of the difference between safety and discriminative abilities fail to look into whether we systematically tend to attribute knowledge in cases where safety and discrimination come apart (2004: chap 2, 64ff.).

presuming a particular view on switching will the question of presentations become particularly pertinent (**§3.4.1**). I will show that an individual's relevant mental states are readily indiscriminable to them in a switching cases given two theoretically significant modes of presentation. In actual switching cases, the relevant states are readily indiscriminable to the individual given a *successive mode of presentation* (**§3.4.2**). However, and possibly more frequently, we consider switching cases as counterfactual scenarios. In that case, the individual will have to rely on some *descriptive mode of presentation* to refer to their counterfactual mental state after switching (**§3.4.3**).

3.4.1 Ready indiscriminability and the conceptual effects of switching

Let us discuss the conceptual effects of switching using the classic switching case as example. There is a controversy among state externalists about the effects that switching will have on the natural kind concept that Oscar expresses with "water".[51]

Some state externalists believe that Oscar loses the set of old concepts once he has been living on Twin Earth for a sufficient period of time (Ludlow 1995; Tye 1998). The assumption is that a conceptual switch is, after some time, fully completed: an individual loses any concept that they acquired in an environment with which he no longer has causal contact. Let us call this the *one concept view*. Proponents of the one concept view have sometimes called it "strong externalism" (Tye 1998: 81) because it would consider an individual's present as well as past external mental states to be fully determined by the individual's present environment after a sufficient period of time. It is a consequence of this view that Oscar, presently on Twin Earth, misremembers events that took place on Earth involving water as involving twin water (Tye 1991: 80ff.).

---

[51] The disagreement on the conceptual effects of switching, evidently, concerns thought contents, not attitudes. However, insofar as mental states are individuated by an attitude as well as a content, different views on the effects of switching on the content thought will affect whether and for what reason individuals are unable to readily discriminate between their mental states in a switching scenario.

Others think that Oscar will continue to have certain semantic ties to Earth even after having lived on Twin Earth for a long time given that Oscar is a normal human adult with a normally functioning memory and conceptual abilities. Oscar remembers many episodes clearly in which he interacted with water: he remembers that he went swimming as a boy or that he used to experiment with distilled water in school, etc. According to those theorists, it is unlikely that the conceptual switch is ever fully completed. However, different theorists have argued for different ways in which the semantic ties to Earth manifest on Twin Earth.

According to some, Oscar's concept WATER broadens into an amalgam concept which we may label as TWIN WATER-WATER that has both water and twin water as its extension as a consequence of the switching (Heal 1998). Let us call this the *amalgam concept view*. Proponents of the amalgam concept view point out that Oscar would mistakenly point out samples of water as well as of twin water as examples of the same natural kind. Oscar is further prone to making false identity judgments after having lived on Twin Earth for a long time ("this is the very stuff I used to experiment with as a boy"). Proponents of the amalgam concept view take those points to suggest that Oscar has too few natural kind concepts in his repertoire. A consequence of the amalgam concept view is that Oscar's natural kind concept WATER has turned into the non-natural kind concept TWIN-WATER-WATER like our actual non-natural kind concept JADE that takes nephrite and jadeite as its extension. It is presumably a consequence of the amalgam concept view that Oscar's memories mistakenly embed the amalgam concept after the switch (although Heal (1998: 109) is indecisive about this).

According to others, Oscar would come to possess the concept WATER as well as the concept TWIN WATER through the switch. He would use each concept depending on whether he is thinking about events that took place on Twin Earth or on Earth (Boghossian 1992; Gibbons 1996; Burge 1998). Let us call this the *two concepts view*. Proponents of the two concepts view tend to emphasise purely historical factors as content-determining whereas proponents of the amalgam-concept view also consider current dispositions of the individual (such as their disposition to classify certain

samples as belonging to a kind) as content-determining. Proponents of the two concepts view also believe that we have independent reasons to think that the function of memory is to rigidly preserve the content of thoughts (see Burge 1993, 1998). Since a majority of theorists defend a historical take on meta-semantics, but also believe that Oscar correctly remembers many past interactions with water, the two concepts view is the most popular view on the conceptual effects of switching among state externalists (Boghossian 1992; Burge 1998).

What concepts Oscar possesses after switching is relevant when we are assessing whether and how some relevant thought contents are readily indiscriminable to him.

Take the one concept view. In the actual context on Twin Earth, Oscar is no longer capable of grasping water-thoughts: solely interacting with twin water has fully replaced any conception of water. To discriminate between two individuals *a* and *b*, I said, is to activate knowledge that *a* and *b* are distinct, under respective modes of presentation. Consider first the case where Oscar's occurrent thought is presented to him as "my current state of thinking that seawater is not drinking water". On the one concept view, Oscar cannot activate knowledge that his occurrent twin water-thought is distinct from his past water-thought because he can no longer grasp a thought with the embedded concept WATER. So, in this case, Oscar is trivially unable to discriminate his actual twin water-thought using the expression "water" from past water-thoughts using the same expression because, Oscar is trivially unable to discriminate anything from something he cannot grasp. Yet, in other cases, Oscar may refer to his past water-thought via some descriptive-demonstrative mode of presentation such as "that thought I had when I went swimming ten years ago", by which he will pick out the water-thought that seawater is not drinking water that he had ten years ago. But, by hypothesis, Oscar is unaware of the fact that he has been switched, so he does not know whether the thought he had when he went swimming ten years ago is distinct from the thought he has now unless he finds out about the switch. So, again, Oscar will not be able to discriminate his current thought from his past thought, although not trivially so.

Take the amalgam concept view next. If Oscar on Twin Earth thinks the thought that he would express by "seawater is not drinking water", then this thought is true if and only if either seawater is not drinking water or twin seawater is not twin drinking water or both cannot be drunk. Presumably, it is a consequence of the amalgam concept view that Oscar's memories mistakenly embed the amalgam concept after the switch. It is definitely a consequence of the amalgam concept view that, were we to continue the switching thought experiment, and imagined that Oscar is switched back to Earth and would continue to live there, his amalgam concept TWIN WATER-WATER would stay the same. Consider the case where, in such a scenario, Oscar is now thinking on Earth what he would express by "seawater is not drinking water". According to the amalgam view, Oscar would thereby express the same thought as he did on Twin Earth (after the first switch) using the same word-form. If the amalgam concept also infiltrates Oscar's memories after the first switch, his (mistaken) memory of thinking that "seawater is not drinking water" on Earth (before any of the switches) would also express the same thought as his thoughts on Twin Earth using the same word-form (after the first switch) as well as his thoughts on Earth using the same word-form (after the second switch). So, on the amalgam concept view Oscar is unable to readily discriminate his relevant thought in the switching case because they are the same thought. Trivially, it is impossible to discriminate anything from itself. In other cases, Oscar may manage to pick out the non-amalgam water-thought he had ten years ago on Earth by virtue of some descriptive or demonstrative mode of presentation. But, again, in virtue of the fact that is unaware of the switch, he will not be able to discriminate his past thought from his current amalgam thought, even if not trivially so.

According to the two concepts view, if Oscar, presently on Twin Earth, believes the twin water-thought that he would express by "seawater is not drinking water", there is a past or counterfactual water-thought that he would express using the same words. He is still capable of grasping water-thoughts because he has not lost his Earthling conceptual abilities. If Oscar is presently able to think twin water- as well as water-thoughts, although in different contexts, the question of presentations becomes pertinent. Is Oscar unable to

readily discriminate his actual twin water-thoughts from his past or counterfactual water-thoughts? It is not obvious that he is. For consider that Oscar has compatibilist access to his present twin water-thought. From this he is capable to infer that he is not presently thinking about water, if "water" is introduced to him as a liquid superficially just like twin water, but actually different. It is clear that given some modes of presentation, Oscar is able to readily discriminate his actual twin water-thought from past water-thoughts. In the next two sub-sections, I will defend two theoretically significant modes of presentation under which the relevant external states of an individual in a switching case are readily indiscriminable to that individual.

3.4.2 Successive mode of presentation

In the classic switching case, Oscar is actually moving back and forth between Earth and Twin Earth, unaware of his condition. Given this set-up, it is natural to think that Oscar is unable to readily discriminate between his particular occurrent twin water-thought and the past water-thought under a successive mode of presentation. We can imagine Oscar thinking a thought that he would express by "seawater is not drinking water" on Earth at t_1 and a thought that he would express by "seawater is not drinking water" on Twin Earth at t_2. Between t_1 and t_2, Oscar has been switched and enough time has passed for him to successfully interact with twin water such that the concept he expresses by "water" now refers to twin water. At t_2, Oscar may refer to his occurrent thought as "my current state of thinking that seawater is not drinking water" and to his past thought as "my earlier state of thinking that 'seawater is not drinking water'".[52] Oscar would seem unable to readily know that these are distinct thoughts. More generally, the proposition is that an individual is unable to readily discriminate between their states in an actual switching cases under successive modes:

[52] One could think that Oscar may be able to discriminate between these thoughts in virtue of their occurrence at different moments in time. But time only accounts for token presentations; states may be presented under the same type of presentation on different occasions.

> *Readily indiscriminable external mental states [SUCCESSIVE]*: S is unable to readily discriminate an occurrent external mental state M_β of S at t_2 from an earlier external mental state M_α of S at t_1 if M_β is presented to S as "my current state of thinking that *p*" and M_α is presented to S as "my earlier state of thinking that '*p*'".

I will argue that individuals are unable to readily discriminate between their states presented to them successively. I will defend the successive mode of presentation against two objections, the *objection from behaviour* and the *objection from memory*.

There are a number of tests one can appeal to in order to see whether an individual is able to readily discriminate between mental states that they are in successively. First, if Oscar is able to readily discriminate between distinct mental states presented to him successively, he should be able to readily notice that a change of his mental states took place. Second, Oscar should be able to readily make reliable judgments about sameness and difference of the relevant mental states. Third, Oscar should be able to react differentially to the change in his mental states. Following Brown (see Brown 2004: 49ff.), I will argue that Oscar fails each of those tests. However, responding to an objection to the last behavioural test, will require qualifying the successive mode of presentation to an individual's *general* as opposed to their specific inability to readily discriminate between their states.

It is part of the set-up of the switching case that Oscar is and remains unaware of the switch. It is further stipulated that the only difference between Earth and Twin Earth is a difference that Oscar is unaware of, namely the difference in the chemical composition of the dominant liquid. It seems clear that Oscar will not be able to notice that his mental states changed at some point after the switch unless he is given more empirical information. Oscar himself would also deny that his environment or some of his mental states have changed in virtue of a switch. Owing to the fact that Oscar is unable to readily notice the change, he is unable to make reliable judgments about the sameness and difference of some of his external mental states. After the switch, Oscar would falsely claim that he is currently thinking the same

thought as he was thinking in the past. Finally, Oscar is not able to readily act differentially to his water-thoughts and his twin water-thoughts. We could reveal everything to Oscar about his situation, except for the moment in which the mental state switch occurs: we may introduce Oscar to the true theory of natural kind externalism and tell him that he will be switched to an environment that is such that a change to the concept that he expresses with "water" will take place. We could ask Oscar to press a button as soon as he starts to express the concept TWIN WATER with "water". It is clear that Oscar will not be able to press the button accurately above chance unless he is given more empirical information about the time of the switch.

Some may object to the last point by insisting that Oscar is capable of acting differentially to his water- and twin water-thoughts even in the original set-up where he remains unaware of the switch. They may point out, for example, that Oscar reacts to his desire to drink some water on Earth by getting some water and he reacts to his desire to drink some twin water on Twin Earth by getting some twin water. Even if Oscar performs exactly the same kind of bodily movement and would himself describe the action he performs in each situation as "getting some water", he would perform two different kinds of actions according to this line of objection, an action of getting some water in one context, and an action of getting some twin water in the other context, responding differentially to distinct mental states. This is the *objection from behaviour* to the successive mode of presentation.

In response to this objection, we can notice that even if we granted that Oscar behaved differentially in those specific situations, it would not be correct to say that Oscar has the *general* ability to act differentially to his water-thoughts and his twin water-thoughts. General abilities, in contrast with specific abilities, are those abilities that individuals have stably, across a number of contexts. Whittle, for example, distinguishes between "what an agent is able to do in a large range of circumstances, and what the agent is able to do now, in some particular circumstances" (Whittle 2010: 2). Oscar is not able to act differentially to his water-thoughts and his twin water-thoughts, respectively, across a large number of contexts. We could ask Oscar to clap if he is thinking a water-thought and to jump if he is thinking a twin water-

thought, or to press different buttons, sing, scream or perform many other differentiating actions in response to his respective thoughts. Oscar will not be able to react differentially above chance because he is unable to reliably judge that he is thinking one thought rather than the other. His ability to react differentially to his water- and twin water-thoughts is limited to those situations that involve interacting with either substance directly and do not require Oscar to be able to reliably judge that he is thinking one thought rather than the other. So, Oscar does not possess the general ability to act differentially to his water- and twin water-thoughts, respectively.

We can adapt the mode of presentation accordingly:

> *Readily indiscriminable external mental states [SUCCESSIVE]**: S is *generally* unable to readily discriminate an occurrent external mental state M_β of S at t_2 from an earlier external mental state M_α of S at t_1 if M_β is presented to S as "my current state of thinking that *p*" and M_α is presented to S as "my earlier state of believing that '*p*'".

Another objection that someone might level against the claim that, in a switching case, Oscar is generally unable to readily discriminate between his water- and twin water-thought given a successive mode of presentation makes use of the point that, necessarily, the successive mode of presentation assumes that Oscar is able to remember at t_2 while being in M_β that he was in M_α at t_1. Recall that proponents of the one concept view believe that Oscar loses his Earthly concepts at t_2, and is therefore unable to remember the external mental states he was in at t_1 (Ludlow 1995; Tye 1998) (this is, most likely, a consequence of the amalgam concept view as well). It is open to proponents of the one concept view to argue that Oscar is in fact generally able to readily discriminate his water-thoughts from his twin water-thoughts, he is simply unable to remember his water-thoughts once he is in a twin water-thought. We know from other examples that such cases are possible. Sometimes we are generally unable to discriminate a certain mental state M_β at t_2 from an earlier mental state M_α at t_1 because at t_2 we are unable to remember M_α at all or with a sufficient level of detail, and not because we are actually generally unable to

readily discriminate M_α from M_β. For example, I may be generally unable to readily discriminate the colour of my carpet from the colour of some curtains when I am at the curtain-shop because I cannot recall the colour of my carpet with sufficient detail. Had I brought a sample piece of the carpet with me, however, I would be able to readily discriminate the two colours (assuming that they are actually distinct). Similarly, proponents of the one concept view may suggest that Oscar is in fact able to readily discriminate his occurrent twin water-thought from his earlier water-thought, he just cannot remember the water-thought once he is in the twin water-thought. This is the *objection from memory* to the successive mode of presentation.

Again following Brown (2004: 56ff.), I will argue that the memory objection fails: a proponent of the memory objection cannot point to any evidence that would suggest that an individual's failure to readily discriminate between their states in a switching case is due to their memory failing. For it is possible that Oscar is unable to readily discriminate between his present water-thoughts and his counterfactual twin water-thoughts both because he is unable to discriminate between his water-thoughts and his twin water thoughts, *and* because he is unable to remember his water-thoughts once he is thinking twin water thoughts. To bolster the view that it is merely in virtue of his failing memory that Oscar is unable to discriminate between his water-thoughts and his twin water-thoughts, one concept theorists have to present evidence that suggests that only Oscar's memory is at fault and not his discriminative abilities.

Again, one can usually run certain tests in order to find out whether a failure to discriminate is due to a memory failing or a lack of discriminative abilities. The most straightforward test is of course to present the two items up for discrimination simultaneously as one could do in the curtain-shop case by bringing a sample piece of the carpet to the curtain-shop. This test is of course not available in the classic switching case: Oscar cannot be on Earth and Twin Earth simultaneously, nor can he be in a water-thought that he would express by "*p*" and a twin water-thought that he would express by "*p*" simultaneously. A slightly more indirect test is to see whether at t_2, an individual is able to provide a discriminating description of the item that cannot be present for

simultaneous comparison. Oscar will of course not be able to provide a discriminating description of his water-thoughts once he is in a twin water-thought because, by hypothesis, Oscar is unaware of any difference between water and twin water, and the corresponding natural kind-thoughts. The memory objection cannot be substantiated.

3.4.3 Descriptive mode of presentation

For the rest of us, inter-planetary travel is reserved to science-fiction: none of us has a past on a different planet. When we consider how our concepts would change if we lived on Twin Earth, we refer to Twin Earth and to our mental states on Twin Earth via some sort of description. The scope of the access problem would be limited if it could only be shown that those individuals who have actually undergone a switch are unable readily know that they are in some particular external state rather than some alternative state.[53] To see whether the problem of access to some of an individual's mental states extends beyond actual switching cases, we will have to see whether individuals are unable to readily discriminate between some occurrent external mental state and some counterfactual external mental state if they refer to the counterfactual mental state via some description.

Let us work with a set-up similar to the situation that we are in when we are first introduced to the Twin Earth scenario. Oscar is on Earth thinking that seawater is not drinking water. Oscar is introduced to the true theory of natural kind externalism. Further, Oscar is told about Twin Earth as a possible planet that does not differ from Earth in anything except for the chemical composition of the dominant liquid. However, an important difference between most of us and Oscar is that Oscar is chemically ignorant: he does not know of the specific chemical composition of water, or of twin water. Oscar reasons that it follows from natural kind externalism that if he was switched unawares to Twin Earth, the concept he expresses with "water" would change

[53] Note though that the risk of being switched must still be relevant otherwise the discrimination condition does not apply. The reality of situations in which we run the risk of being switched will be discussed in §3.5.

to express a different concept after some time. Oscar may refer to his occurrent thought that seawater is not drinking water indexically as "my current state of thinking that seawater is not drinking water". Is there a description for the counterfactual twin water-thought that Oscar's counterpart on Twin Earth would also express with "my current state of thinking that seawater is not drinking water" that is such that Oscar is unable to readily discriminate his current water-state from the counterfactual twin water-thought?

A natural proposition is that Oscar cannot readily discriminate his occurrent water-thought on Earth presented to him as "my current state of thinking that seawater is not drinking water" from his counterfactual twin water-thought presented to him as "my counterpart's state of thinking that 'seawater is not drinking water'". Parrott and Gomes, for example, propose a descriptive mode of presentation along those lines. They write, "in other words, it is not possible for an individual to know through introspection that her current thought is not one of the twin water-thoughts" (Parrott and Gomes 2021: 327). Similarly, Sawyer writes about a subject Susan who is in the classic switching case: "Susan's thought is a member of a set of epistemic counterparts, she cannot distinguish it *a priori* from any other member of that set. Hence, she cannot know *a priori* it is that thought she has, as opposed to any one of its epistemic counterparts" (1999: 363).[54] More generally, the proposition is that an individual's thoughts are readily indiscriminable to them in a switching case given the following descriptive mode of presentation:

> *Readily indiscriminable external mental state [DESCRIPTIVE]*: S is unable to readily discriminate an occurrent external mental state M_α of S presented to S as "my current state of thinking that *p*" from a counterfactual external mental state M_β of S presented to S as "my counterpart's state of thinking that '*p*'".

[54] Parrott and Gomes, as well as Sawyer, take introspective ("special" in my terminology) access, not ready access, to be the relevant kind of access in the discrimination argument.

But things are not so simple. Whether Oscar is unable to discriminate his current water-thought from his counterpart's twin water-thought depends on how Twin Earth and his counterpart on Twin Earth are introduced to Oscar. In the current set-up, Oscar has been introduced to Twin Earth as a possible planet just like Earth except for the chemical composition of the dominant liquid. Oscar will conceive of his counterpart accordingly as someone just like him on a planet that is just like Earth except that the dominant liquid is chemically different on that planet. Is Oscar unable to readily discriminate his current water-thought from his counterfactual's twin water-thoughts? We may imagine Oscar asking the following questions:

> "How do I know that this experience I am having is not a counterfactual experience, which I don't have, but could have had, should I have been brought up in a different environment? Or how do I know that the experience I am having of this stuff, is not some different experience some other bloke is having of another stuff at another part of the universe?" (See Farkas 2008: 107).

It is not very difficult for Oscar to answer these questions. Oscar will be able to readily, namely *conceptually*, discriminate his current water-thoughts from those thoughts that his counterpart is thinking when his counterpart uses the expression "water" to think about a *different* liquid on *another* planet. Some descriptions will make Oscar's current states *trivially* readily discriminable from counterfactual states expressed under the guise of the same sentence.

We have been supposing, as in the original Twin Earth story, that Oscar is chemically ignorant: Oscar does not know that water is H_2O and he does not know that twin water is XYZ. Now let us say that Twin Earth is introduced to Oscar as a planet where the dominant liquid that people on Twin Earth call "water" is XYZ, i.e. its molecular structure is not H_2O. Then it would seem possible that Oscar knows that he is thinking that seawater is not drinking water without knowing that he is not thereby thinking the thought that his counterpart would express by "seawater is not drinking water". This is because, in this case, Oscar does not know whether the concept he expresses with "water" is the same as the concept that his counterpart expresses with "water". If all Oscar knows about twin water is that it is not H_2O, and he does

not know that water is H_2O, then Oscar does not know whether "water" and "twin water" are synonyms, or whether the concept WATER is identical to the concept TWIN WATER. In that case, Oscar does not know whether when he utters "seawater is not drinking water", and when his counterpart utters the same sentence, they express the same proposition.

But here is another piece of chemical knowledge that Oscar lacks: he also does not know that twin water is not $C_2H_4O_2$. But as a matter of fact drinking acetic acid ($C_2H_4O_2$) is a very different experience from drinking water (for the example, see Farkas 2008: 108). Nevertheless, if "counterpart" is introduced to Oscar as "someone just like you who lives in an environment where water is $C_2H_4O_2$", Oscar will also not know whether, when he utters "seawater is not drinking water" and when his counterpart utters the same sentence, they express the same proposition. Or consider yet another description by which "counterpart" is introduced to Oscar: "someone just like you who lives in an environment where water is the liquid that appeared to you in a dream ten years ago". Yet, what Oscar does not remember is that the liquid that appeared to him in a dream ten years ago was some horrid, yellow fluid easily distinguished from water. What this shows is that there is probably *some* description under which *any* two particulars would not be readily discriminable however obviously distinct the two particulars actually are: for example, Fred's favourite fruit is indiscriminable to you from John's favourite fruit, if introduced to you under those modes of presentation even if one is an apple and the other an orange (of course, assuming you do not know what their respective favourite fruit is). So the fact that Oscar is unable to readily discriminate between his current thought and the thought his counterpart would express using the same sentence does not show that state externalism constitutes a limitation on our self-knowledge in a way that state internalism does not. If Oscar were switched to the environment in which the substance that runs in rivers and comes out of showers is acetic acid, he would quickly notice that that which people in that environment call "water" is distinct from what people on Earth call "water". State internalists will agree that "water" expresses two distinct concepts on Earth and in the acidic environment. So, some descriptions are such that they make Oscar's current thought *irrelevantly*

indiscriminable to Oscar from the thought his counterpart would express in the same way.

What descriptions are such that they would make Oscar's current states *relevantly* indiscriminable from his counterpart's states? Relevancy is a context-sensitive notion and the present context is to see whether, assuming state externalism, distinct external states may be readily indiscriminable given certain descriptions to an individual in switching cases. Switching cases, crucially, are relevant twin cases and twin cases are the central cases in the debate between state internalists and externalists. State internalists believe that an individual in and out of a twin case is in the same mental state whereas state externalists hold that they may be in distinct mental states. A description that would make Oscar's current thought relevantly indiscriminable from his counterpart's thought is therefore a description that captures the fault line between state internalists and externalists. The description must therefore be such that, plausibly, if Oscar's current thought is readily indiscriminable from his counterpart's thought under that description, state internalists will claim that Oscar is thinking the same thought in the present and in the counterfactual environment whereas state externalists will claim that he may be thinking distinct thoughts.

As Farkas points out, the most direct way to spoil a twin case such that it no longer divides state internalists and externalists is (Farkas 2008: 81), to render the cases readily discriminable. Putnam's original Twin Earth case would not work if twin water was bitter or dark, or if it was acetic acid for that matter. As mentioned, in such a case, it would not be hard for state internalists to explain why Twin Oscar's thoughts about what he calls "water" differ from Oscar's about what he calls "water". The appropriate description by which "counterpart" (and, relatedly, "twin water" and "Twin Earth") is introduced to Oscar must therefore guarantee that the environment in which the counterpart lives, i.e. Twin Earth, differs from Earth only with respect to a property hidden from (chemically ignorant) Oscar, namely the molecular structure of the dominant liquid. If "counterpart" is introduced to Oscar as "someone just like you who lives in an environment *that is readily indiscriminable from your environment* where what they call 'water' is not H_2O", Oscar is unable to

readily discriminate his current water-thoughts from his counterpart's twin water-thoughts that his counterpart would express under the guise of the same sentences.

We can modify the above definition of the descriptive mode of presentation accordingly:

> *Readily indiscriminable external mental state [DESCRIPTIVE]**: S is unable to readily discriminate an occurrent external mental state M_α of S presented to S as "my current state of thinking that *p*" from a counterfactual external mental state M_β of S presented to S as "my counterpart's state of thinking that '*p*'" where "counterpart" is defined as S's duplicate who lives in an environment of which S does not readily know that it is not S's environment.

In other words, Oscar is not able to readily discriminate his occurrent states from the mental states he would be having on Twin Earth because he does not readily know that Earth is not Twin Earth. This is precisely a set-up that we would expect to divide state internalists and externalists, i.e. where state internalists will claim that, if the only difference between the two environments is hidden to Oscar in this way, Oscar's mental states cannot have changed in virtue of the switch.

Coming back to the discrimination argument, we are assuming the truth of premise (1) for the sake of discussing the *Don't-worry-strategy*. Premise (2) is true by hypothesis about switching cases. In this section, I defended two versions of premise (3) of the discrimination argument:

> (3) In a switching case, S is generally unable to readily discriminate the case α in which S is in M_α presented to S as "my current state of thinking that *p*" from the relevant alternative case β in which S is in M_β presented to S *either* as "my earlier state of thinking that '*p*'" *or* as "my counterpart's state of thinking that '*p*'" where "counterpart" is defined as S's duplicate who lives in an environment of which S does not readily know that it is not S's environment.

The discrimination argument against ready access is sound: in switching cases, individuals are unable to readily know that they are in a particular occurrent external state rather than a particular alternative external state.

If Oscar's and Twin Oscar's mental states are readily indiscriminable given the appropriate successive and descriptive modes of presentation in switching cases, then so are Gary's and Barry's mental states. Further, switching cases can be formulated for any state that is externally-individuated according to state externalists. So, *all* of Gary's and Barry's external states are readily indiscriminable to them assuming an appropriate switching scenario. Further, by hypothesis, Gary and Barry are in the same purely internal states. In other words, in an appropriate switching scenario, Gary and Barry are "indiscriminables" in the way defined in §2.4: all of their mental states are readily indiscriminable to them either in virtue of the appropriate modes of presentation defended in this section or in virtue of the reflexivity of indiscriminability.

3.5 The Objection from Relevance

In switching cases, the alternative external state is relevant by hypothesis. But that does not mean that it is a relevant alternative for *us* that we are in some alternative, indiscriminable mental state. A number of compatibilists who embrace the *Don't-worry-strategy* argue that we do not have to worry about the discrimination argument because switching cases are not normally relevant to us (see Warfield 1997; Sawyer 1999; Brown 2004). In most cases, they say, we are able to readily discriminate our current external state from any relevant alternative state. If that is correct, we would be able to readily know that we are in a particular state by the light of the discrimination condition on knowledge. Sawyer, for example, argues that, in most cases an individual Susan (briefly mentioned above) will be able to readily discriminate her belief that water is wet from "various other propositional mental events she may

have, such as the belief that that's Ortcutt, the fear that her car won't start, the desire for a gin and tonic, and so on" (Sawyer 1999: 370). At the same time, Susan may be unable to discriminate her belief that water is wet from the epistemic counterpart thought that twin water is wet, but, Sawyer continues, "the mere possibility that there be such duplicate thoughts need not (…) defeat her claim to know, on any particular occasion when she has formed the true belief that she thinks that water is wet on the basis of exercizing that recognitional capacity, that she believes that water is wet" (*ibid.*). With respect to certain external states, proponents of the *Don't-worry-strategy* even suggest that we *could* not normally be at risk of being switched and be in the states that we are actually in.

So far, I have focused on the classic switching case of Oscar ignorantly being switched between Earth and Twin Earth. If this is the most relevant example of a switching case, I agree that we would not need to be very worried. However, even if switching cases turn out to be rare with respect to natural kind-thoughts, natural kind externalism is just one type of state externalism. Switching may be a more frequent occurrence with respect to other externally individuated mental states. A proper response to the objection from relevance requires looking at the different forms of state externalism and their propensity for switching. For each of the main types of state externalism, I'll look into the likelihood of actual switching cases: **§3.5.1** looks into switching cases involving natural kind-thoughts; **§3.5.2** looks into switching cases involving socially individuated thoughts; **§3.5.3** looks into switching cases involving singular thoughts, and, finally, **§3.5.4** looks into switching cases involving factive states. I will argue that, if we consider all kinds of externally-individuated mental states, switching cases are sometimes relevant to us.[55]

3.5.1 Natural kind externalism and the relevance of switching

[55] I do not discuss switching cases with respect to phenomenal properties in this section because they are not relevant to the discrimination argument about ready access to our external *states*.

Switching of natural kind-thoughts occurs in those cases in which the environment or parts of the environment of an individual are substituted for another that contains a readily indiscriminable natural kind in such a way that the individual remains oblivious to the change unless they receive further empirical information.

Note, however, that while there are many natural kinds that are readily indiscriminable to the layperson— think of elm trees and beech trees, crows and rooks, walleyes and saugers—it is not enough to yield a natural kind-state switch in an individual if that individual moves to an area where, unbeknownst to them, a distinct but (to them) not readily discriminable natural kind dominates. Consider the following scenario. An individual lives in an area where elm trees are the dominant kind of tree, they acquire the concept ELM TREE and successfully refer with "elm tree" to elm trees. They then move to a different part of the country where, unbeknownst to them, only beech trees grow. They do not possess the concept BEECH TREE and, mistaking the beech trees for elm trees, use the expression "elm tree" to speak about the local trees. Even if they stay on to live in that area for years, using the expression "elm tree" to speak about the beech trees in their garden, never interact with actual elm trees, maybe even teach their children to call the beech trees "elm trees", the concept that they express by "elm tree" will not at some point be BEECH TREE. They and their family will simply have said many false things about elm trees. This is because they had causal contact with elm trees, acquired the concept ELM TREE in interaction with those and presumably still have contact with people who causally interact with elm trees and use "elm tree" to refer to elm trees. All of these factors will ensure that their expression "elm tree" will continue to refer to elm trees.

We may imagine a more radical variant of the same story that would presumably support a different verdict. Take the same set-up but add that the individual cuts all ties to their old life when moving to the new area, as well as any ties to people who interact with elm trees and call them "elm trees", and all of their interactions are henceforth with beech trees. In such a case, it may be that, after sufficient time has passed, the concept that the individual (and their family) express with "elm tree" will be BEECH TREE. But it is clear that

these qualifications do not make it any more likely that a natural kind-thought switch is a common occurrence for the rest of us.

It is likely that the most common examples of natural kind-thought switching are cases usually discussed as examples for socially individuated thoughts. Take someone who moves between Britain and the Unites States. They may not realise that "chicory" designates a slightly different vegetable in British and American English (see Ludlow 1997 for the example), namely *cichorium endivia* is in the US and *cichorium intybus* in the UK. Let us add, less realistically, that this is because only *cichorium endivia* is grown in the US, whereas only *cichorium intybus* is grown in the UK. Now consider an American who has just arrived in Britain and sincerely says "Chicory$_{AM}$ is rich in vitamin A" thereby expressing a belief about *cichorium endivia*. After having lived in Britain for enough time, successfully interacting solely with *cichorium intybus* and having become a part of the new linguistic community, this belief will presumably change to express a belief about *cichorium intybus*. In such a case, given the discrimination argument, the American would fail to readily know that they believe that chicory$_{UK}$ is rich in vitamin A as they would not be able to readily discriminate that belief from their previous belief that chicory$_{AM}$ is in rich in vitamin A, without receiving more empirical information about their natural and linguistic environment. But, even counting such cases, switching cases involving natural kind thoughts will be rare.

3.5.2 Social externalism and the relevance of switching

There are a number of terms that have a slightly different meanings in different linguistic communities. Beside the mentioned "chicory", "jelly", "pants", "wasp" are further examples of expressions that are defined in slightly different ways in American and British English, for example. For any of these terms, it is possible to imagine a case of someone moving from the American to the British linguistic community, or *vice versa*, who remains oblivious to the differences between them. As Ludlow and others have pointed out switching cases may also occur when people move across smaller social

groups that have their own internal conventions for the use of certain terms without realising it (Ludlow 1997: 46-49; Goldberg 1999). Philosophers, for instance, use terms like "realist" and "pragmatist" in a specialised way that differs from general public usage but not obviously so. We could imagine a switching case of someone who starts to socialise with philosophers more, and defers to philosophers over some terminological questions. They may acquire a new concept, Pragmatist$_{PHIL}$, that refers to pragmatists in the sense of the philosophical school of thought. There is some semantic overlap to the concept as it is standardly used, for example that pragmatists in the standard and in the philosophical sense do not believe in principles. In virtue of the overlap, they may remain oblivious to the fact that they acquired a new concept. In some contexts, they may then express the concept Pragmatist$_{PHIL}$ when they use the word "pragmatist" without realising it.

However, the smaller and the less separated the linguistic group is from a larger linguistic group, the more doubtful it is that an individual would acquire a new distinct concept instead of learning new applications of their old concept. Here is a switching case involving socially-individuated thoughts that most clearly avoids that worry.

Olivia who grew up bilingual in English and French: she gets to spend the summers in England, and the winters in France (she cannot complain). In French, she acquires the word "prune": she believes that "prunes" are purple fruits that grow on trees, have fairly big pits, and that she has had them on cakes. Back in England, she also acquires the word "prune": she believes that they are purple fruits with big pits, that in dried form are perfect for hikes, etc. She also believes that "prune" means the same in French and in English, the only difference being in pronunciation. She is wrong: "une prune" in French is a plum, and "a prune" in English, a dried plum, is "un pruneau" in French. So Olivia has some false beliefs about the application of each concept; for example, she would agree to the false claim that "prunes grow on trees". Nevertheless, it would seem that Olivia has enough understanding of the concept in each language to acquire and use it successfully: she defers to experts of French when in France, and to experts of English when in England; she has many true beliefs about the application of the concepts in both

languages; and she makes herself understood successfully often enough. In virtue of the fact that the concepts are part of different languages, clearly used in different contexts and regulated by different experts, it seems clear that Olivia possesses two distinct concepts, namely $PRUNE_{Engl}$ and $PRUNE_{French}$. If that is correct, Olivia is a competent user of two distinct concepts each of which she partially understands. Since Olivia believes that "prune" in English and French are synonyms, she takes herself to express the same belief when she says "J'adore les tartes aux prunes" and "I love prune tarts". In fact, she expresses a belief about plums, and a belief about prunes. In a particular case in which Olivia is thinking "I love prune tarts", she would fail to readily know that she is thinking about prunes because she cannot readily discriminate this belief from a belief about plums.

What these examples show is that compared to switching cases with respect to natural kind thoughts, switching cases with respect to socially individuated thought occur more regularly. Still, they are hardly something that we would expect to undergo on a daily basis. Indeed, switching cases may happen on a handful of occasions during a lifetime.

In fact, Brown argues that such switching cases could not occur so commonly as to become the norm (2004: Chap.4). Brown summarises the ingredients of switching cases with respect to socially individuated thoughts as follows:

(i) S is switched between two linguistic communities that share a single word but define it slightly differently;
(ii) S is ignorant of the linguistic difference;
(iii) S is a competent speaker of both languages and of the target word in both languages (2004: 140).

If those features are satisfied, it is plausible that some of S's mental states may switch in a way that S is unable to notice without further empirical information. Take Olivia. We consider her to possess two concepts—$PRUNE_{Engl}$ and $PRUNE_{French}$—only if she competently employs each of them in most contexts. Had Olivia been a monolingual speaker of French, she would

not count as expressing a belief about prunes even as she exceptionally utters the English sentence "I love prune tarts". But to be a competent user of two distinct concepts without noticing it, the difference in meaning can only be marginal.

Brown argues that if switching became the norm, conditions (i) and (ii) would be undermined. This is because, if two linguistic communities interact more regularly, words that previously had different meanings will settle on a single meaning both by use and by the fact that speakers would defer to common linguistic experts. To the extent that differences in meaning survive, it will be less likely for people not to know about them if their linguistic communities are in constant exchange. In other words, plausibly, if exchanges between two linguistic communities become so frequent as to become the norm, the languages become one. But within a language, there is a tendency to avoid ambiguities: differences in meaning will usually exert pressure towards being manifested in a difference in terms.

I agree with Brown's argument for the conclusion that switching cases could not be so common as to become the norm. But Brown wants to draw the stronger conclusion that, therefore, switching cases involving socially individuated thought could not *normally* be relevant to us. Brown casually switches between both formulations. She writes: "It is no accident that conditions (i)–(iii) are not normally jointly met. If slow switching became the norm, this would undermine…" (2004: 141). But less is required for an event to happen normally than for it to constitute the norm.

When does something normally happen? In the present context, and as suggested by Brown's implicit equivocation, we are interested in a frequency-related notion of normality.[56] We may distinguish, roughly, between four categories of frequency: a certain event may never take place, or it may happen rarely, or sometimes or frequently. "Normally" as it is standardly used in English is, as "usually", a generic adverb of quantification that contrasts with the unique occurrence of an event but is appropriately used for a range of regular occurrences (see Loets 2022: §3, for discussion). In contrast, for some

[56] Brown's slipping back and forth is not surprising given different uses of "normally" in English (see Bear and Knobe 2017).

event to constitute the norm, it has to occur in the majority of cases. Certain events happen frequently without being the norm: changes of government happen frequently in the UK, but they are not the norm. We cannot conclude from the fact that switching cases involving socially individuated thought could not be the norm that they could not normally be relevant to us. As I said above, given which conditions have to be jointly satisfied for a switching case involving socially individuated thoughts to occur, they are likely going to be rare occurrences. Nevertheless, and it is important to stress that going forward, in order to show that switching cases are normally relevant to us, one does not need to show that switching cases are or could be the norm.

3.5.3 Singular externalism and the relevance of switching

According to Russellian singular externalism, states of thinking demonstrative thoughts depend for their content on the individual's relations to their immediate environment. Here are examples of switching cases involving demonstrative thoughts as they may happen to us.

Say, you are in a sushi restaurant sitting at the sushi conveyor belt considering which roll to take. You think that that crispy roll (cr) looks good but it is already too far down to take. When a crispy roll is riding in front of you the next time, you believe that it is the same roll and believe that that roll looks good. To you, it will seem that you are thinking the same thought as five minutes ago.[57] But you missed that someone further down the counter took the previous roll and it was replaced by another crispy roll too similar to you to be able to readily discriminate it from the previous one. According to Russellian singular externalism, your demonstrative belief "that roll looks good" will have unwittingly switched to express a different proposition.

Or imagine that you are driving on a packed highway and think that that car (c) in the lane next to you is too large to be moral. Your attention gets

[57] I am assuming an content-internalist friendly, anti-temporalist conception of propositions according to which the same proposition can be expressed at different moments (see, e.g., Salmon 1989). Were the time part of the proposition, the individual would not be thinking the same thought in this case by that fact alone.

diverted for a moment when a car of the same kind pulls in in front of you. You believe that it is the same car that has changed lanes. Again, you would take yourself to think the same thought as five minutes ago if you thought that that car is too large to be moral. But in fact you lost track of the previous car and it is another car of the same kind in front you. Again, according to Russellian singular externalism, your demonstrative belief will have unwittingly switched.

Or you may be looking at flies as they buzz around your head, and think that that fly (f) is fat. Given the number of similar flies in the swarm, it is a relevant alternative that you are instead looking at another fly and think a distinct thought about that fly without realising it.

What these cases show is that, in order for switching cases involving singular thoughts to occur, an object does not need to have a perfect twin. What it takes is that an individual loses track of the object of their thinking momentarily and another object, too similar to be readily discriminable to that individual in that context, takes its place as the object of a singular thought. Still circumstances will have to conspire against the individual in a particular way, with them losing track of an object that is swiftly replaced by a readily indiscriminable object. While such switching cases are likely to occur more commonly than switching cases involving natural kind-thoughts or socially individuated thoughts, they will not happen frequently.

Neo-Fregeans have explicitly designed their version of singular externalism to prevent switching cases from occurring. Neo-Fregean views connect the ability to think perceptual demonstrative thoughts with the ability to track objects over time (Evans 1982; Campbell 1987). According to neo-Fregeans, the individual will suffer an illusion of a demonstrative thought when they lost track of the object (Campbell 1987: 285). If you observe someone walking down the street thinking that that person has an energetic walk, but in fact lose track of the person and think of more than one person that they have an energetic walk, you will, according to these views, suffer an illusion of thought. In such a case, according to these views, you would not think two distinct yet indiscriminable demonstrative thoughts but you would not think a demonstrative thought at all.

However, in order to allow individuals to think of objects they have lost track of, neo-Fregeans usually weaken the tracking requirement and only demand that the individual has tracked the object for a sufficient period of time. After that period of time, they can think of the object in virtue of memory links (see Evans 1982:195–196). But if that is permitted, switching cases are possible. An individual may mistakenly believe of two different objects that they are the same, namely of the object that they are currently tracking, and of a distinct and to them not readily discriminable object that they were tracking for a sufficient amount of time and now think of via memory links. With this weakening in place, the person confusing different people with energetic walks, for example, does not suffer an illusion of thought but thinks distinct yet not readily discriminable perceptual demonstrative thoughts.

3.5.4 Attitude externalism and the relevance of switching

Finally, let us consider switching cases for attitude externalism. Disjunctivists about perceptual beliefs have often developed their views in reliance on the thought experiment of a perfect hallucination (for discussion see Martin 2004, 2006; Farkas 2008). A perfect hallucination is often introduced in something like the following fashion:

> "Suppose I now see a teacup in front of me. Would it not be possible that everything seems the same, and yet the teacup I take myself to be perceiving is not there? Would it not be possible to have a hallucination that was specially indistinguishable from my present experience?" (Farkas 2008: 207).

It is possible, in principle, for a veridical perception to be replaced by a not readily distinguishable hallucination or illusion. In such a case, disjunctivism and anti-conjunctivism predict that the perceiver will undergo a factive-state switching: I will have gone from seeing that there is teacup in front of me to merely seeming to see that there is a teacup in front of me. But

while hallucinations occur, they occur rarely; and it seems doubtful that they are ever perfect.

The more interesting cases of factive-states switching occur if we focus on factive states more generally, and the most general factive attitude, i.e. knowledge, specifically. To use an example of a knowledge switching case owing to Smith (2017: 110), say that this morning Smith knew that his trash bin was on the sidewalk, having put it there himself. His belief persisted until, some hours later, he saw it by the gate. His knowledge, however, will have been lost some time prior to this when his neighbour, seeing that the rubbish had been collected, wheeled the trash bin in. In that moment, Smith's mental state will have switched from a case of knowing that his trash bin sits on the sidewalk to merely believing that it does. But after the switch, Smith is not able to readily discriminate his current state of merely believing that his trash bin sits on the sidewalk from his previous knowledge-state without getting further empirical information (i.e. discovering that it has been wheeled in).[58]

Switching cases involving knowledge will be very frequent: whenever someone loses a piece of knowledge without realising it. Knowledge switching cases will be caused by small changes to our everyday environment that we are unaware of as well as by global changes that take some time to reach us. In a globally connected world of information, we know things about many people and states of affairs physically remote from us whose changes will not be readily accessible to us but will affect what we know.

In conclusion, we can note the following on the respective relevance of switching to us. We will rarely move between different linguistic communities without realising it and think or risk thinking a thought that is readily indiscriminable without realising it; and very rarely, some of our natural kind thoughts will unwittingly refer to a different natural kind. But we will sometimes be unable to readily discriminate between different objects or people and think or risk thinking readily indiscriminable, distinct demonstrative thoughts about them. And our factive states will frequently

[58] Note that even compatibilists about factive states will agree that the mental state switch is not readily discriminable to the individual in cases such as this as they do not defend negative introspection of factive states (see Pritchard 2012; Das and Salow 2016).

switch to non-factive states without our knowledge. Taken together, these cases form a non-negligible package of switching cases that are relevant to us.

3.6 Conclusion

In this chapter, I defended the cogency of the discrimination argument against proponents of the *Don't-worry-strategy*. According to the *Don't-worry-strategy*, we do not have to worry about the discrimination argument against ready access even if the discrimination condition applies to self-knowledge because the argument is argued to fail someplace else. Against the *Don't-worry-strategy*, I showed that individuals are unable to readily discriminate an occurrent external state from a relevant alternative state in switching circumstances given certain successive and descriptive modes of presentation. Further, such switching conditions are sometimes relevant to us. We have to worry about the discrimination argument because, if state externalism is correct, we will sometimes not be able to readily know that we are in a particular external state rather than a particular distinct external state unless we change our epistemic situation significantly by receiving further evidence about our immediate or non-immediate, linguistic or natural environment.

Since the discrimination argument against discriminatory *ready* access is sound, so is the discrimination argument against discriminatory *special* access: in switching cases, individuals are unable to specially know that they are in a particular external mental state rather than a particular distinct state.

The success of a compatibilist response to the discrimination argument fully depends on the success of the *Even-if-strategy*. I pointed out that compatibilists do not have reasons to be too optimistic about the prospects of the *Even-if-strategy*. In case it does not succeed and compatibilist accounts of self-knowledge do not hold up to scrutiny, it would have been established that state externalism is incompatible with ready access to one's external states *tout court* in switching cases, some of which are regularly relevant to us.

What if anything follows from the soundness and relevance of the discrimination argument for hybrid access internalist views? If we sometimes do not know that we are in a particular external state rather than some relevant alternative state, this is likely to be a problem for views that consider justification to be function of those facts to which an individual has ready access. It may be, for example, that hybrid access internalist positions are committed to counterintuitive views on what individuals in switching cases have justification to believe. Whether that is the case, will have to wait for future research. In the last chapter of this thesis, I will argue that we can use the results of this chapter to defend a revisionary definition of the debate between state internalists and state externalists.

Chapter 4

4 - The Discrimination Definition of the Internalist/Externalist Debate about Mind

4.1 Defining the Internalist/Externalist Debate about Mind

In the last chapter, I defended the claim that, on any form of state externalism, individuals are unable to know that they are in a given occurrent external mental state rather than a particular alternative external mental state (assuming the appropriate successive or descriptive modes of presentation) in switching cases without significantly changing their epistemic situation. I also said that in switching cases, an individual experiences or is at risk of experiencing "*mental state switching"* (hereafter, "switching"), where some of their external mental states change and the individual is not in a position to register the change without significantly changing their epistemic situation.

In this chapter, I will defend the claim that this is *the* defining feature of state externalism: a view is state externalist if and only if it allows for switching. The definition of state externalism that I will put forward, reads roughly as follows:

> *Discrimination definition of state externalism*: a theory is state externalist if and only if, according to it, there is a case α that is such that S is unable to readily discriminate the case α in which S is in M_α presented to S as "my current state of thinking that *p*" from an alternative case β in which S is in M_β presented to S either as "my earlier state of thinking that '*p*'" or as "my counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which S does not readily know that it is not S's world), and $M_\alpha \neq M_\beta$.[59]

I will define state internalism as the negation of state externalism: a view is state internalist if and only if the possibility of switching is ruled out. What state internalists and externalists disagree about according to the definition of their debate that I will defend is, roughly, whether an individual's

[59] "Roughly", because, as we will see, a crucial specification is needed of the way in which an individual is unable to readily discriminate between their states, namely *categorically*.

mental states may vary independently of the individual's capacity for ready discrimination.

An individual's capacity for ready discrimination is an epistemic capacity. However, in chapter 1, state externalism was defined as follows:

> *State externalism*: An individual's mental states do not exclusively supervene on the individual's physical or phenomenal states.

This standard definition of the state internalist/externalist debate is metaphysical: the disagreement between state internalists and state externalists is taken to concern the appropriate supervenience base for mental states.

In this chapter, I will argue that standard metaphysical definitions of the state internalist/externalist debate should be rejected in favour of an epistemic definition of their disagreement in terms of certain discriminative capacities of the individual. I am not the first to think that the debate between state internalists and externalists should be defined in epistemic instead of metaphysical terms. In a recent paper, Parrott and Gomes argue that state internalists/externalists "dispute about whether one's mental states can vary independently of certain introspective capacities", and more specifically the "capacity for introspective discrimination" (2021: 331/315). Before Parrott and Gomes, Katalin Farkas raised issues with the metaphysical definition of the state internalist/externalist debate in terms of an individual's *physical states* and stressed that state internalism and state externalism are, at bottom, views about thinkers' epistemic relations to their thoughts (Farkas 2003, 2008). Specifically, Farkas suggests that internalism about thought contents is the thesis that "facts individuate mental contents only insofar as they make a difference to the way things appear to us. This means that any difference in the content of thoughts should be distinguishable from the subject's point of view and hence remains within the reach of privileged access" (Farkas 2003: 203). Farkas takes these considerations to support a metaphysical definition of the state internalist/externalist debate by reference to an individual's *phenomenal states*. While I disagree with the metaphysical conclusions Farkas draws, I

very much agree with the view of the state internalist/externalist debate that motivates them.[60]

As we will see, Parrott and Gomes' epistemic definition of the debate goes some way towards the correct account. They move in the right direction regarding how to solve the challenges that an epistemic definition of the debate faces, but they miss a crucial distinction which results in objections to their account that they cannot resolve. The distinction they miss is that between the definition of a debate on the one hand, and substantial accounts within the debate on the other.

A definition of a debate and substantial accounts within the debate are distinct theoretical items: they serve different theoretical purposes and need to satisfy different desiderata. The definition identifies what the parties in the debate disagree about most fundamentally. It identifies what is common to a type of theory, i.e. what it *means* to put forward a theory of a certain type, such as an externalist theory about a mental phenomenon. Substantial accounts, in contrast, are complete, explanatorily powerful theories. They start off from certain substantial assumptions that will not be shared by every theory even of the same type. A good debate definition is compatible with a number of substantial accounts extant on each side of the disagreement, distinguished by the further negotiable assumptions they embrace.

Applying the distinction between the definition of the debate and substantial accounts within the debate, I will show that the state internalist/ externalist debate must be defined epistemically along the lines of the above discrimination definition, but that substantial accounts within the debate are metaphysically committing views. Certain state internalist substantial accounts will, for example, take an individual's mental states to supervene on the individual's total physical or phenomenal states. Once the distinction between the definition of the debate and substantial accounts within the debate is applied, an epistemic definition of the state internalist/externalist debate can be successfully defended against objections. To stress the contrast between

---

[60] As mentioned, Farkas' view concerns the internalist/externalist debate on thought contents, as is the case for many of the arguments and views that will be discussed in this chapter. In most cases, the views can be adapted to concern internalism and externalism about mental states without loss. In other cases, the issue will be explicitly discussed.

metaphysical and epistemic definitions and accounts that is at issue in this chapter, I will speak, as is more customary, of the "epistemic" instead of the "discrimination" definition for the rest of the chapter.

Before starting, let me respond to a motivational worry one may have about this project: why care about the definition of a debate instead of looking at the substantial accounts and finding out which is true? While defining a debate is not a precondition on making substantial progress on individual theories, being able to define a debate undoubtedly comes with a number of advantages.

As I said, the definition of a debate identifies what the main issue is that opposing sides disagree about. So it also clarifies which issues are derivative. Brie Gertler, for example, complains that state internalist and externalist positions are associated with a "shifting set of claims encompassing a heterogeneous array of topics", topics that include the organism's contribution to the nature of an individual's mental states, links between the individual and their community, and the relations between phenomenal character and intentional content (Gertler 2012: 51). The right definition of the state internalist/externalist debate should say which of these issues is the fundamental point of disagreement between internalists and externalists about mind.

I also said that a definition identifies the defining feature of a type of position in the debate. So a definition will function as a meta-theoretical tool. A good definition provides the basis for formulating a well-motivated taxonomy of positions in the debate.

Finally, in identifying the defining feature of a type of position, a definition helps to identify what are the analytic implications of a position and what are negotiable further assumptions particular views may make. Depending on how far-reaching the consequences of a debate are for further philosophical issues, it may be particularly important to have a clear view on the analytic implications of a certain view. Again, according to Gertler, the lack of a clear definition of the state internalist/externalist debate is especially worrisome, because the debate touches on so many other questions in the philosophy of language, epistemology and philosophy of mind, such as: 'do

thinkers generally have special access to their mental states?' or 'does the meaning of an individual's utterance correspond to what the speaker understands of their utterance?' (see Gertler 2012: 52). Disambiguating the disagreement between state internalists and externalists promises to promote progress on a spectrum of philosophical questions.

We can read off from the advantages of a successful definition of a debate what the desiderata on a good definition are. A good definition

- *identifies* the best candidate for a substantive point of disagreement;
- *sorts* positions in a theoretically motivated and fruitful way;
- and *predicts* the analytic implications of positions.

In the next section, **§4.2**, I will show that every one of the existing definitions of the state internalist/externalist debate fails to satisfy some of these desiderata. I distinguish between three main definitions: the *spatiophysical definition*, the *phenomenal definition* and the *epistemic definition*. The first two are metaphysical definitions, the latter is epistemic. As we will see, the issue with the metaphysical definitions is that they misclassify views as well as make the wrong predictions about the views' implications. The epistemic definition, as we will see, does not misclassify views and makes plausible predictions. It results, however, in an extremely implausible statement of state internalism.

Taking the debate in view, some consider the prospects of defining the state internalist/externalist debate with pessimism. Gertler writes: "The sense that there is a substantive, defining commitment of externalism or internalism —even one that is vague or underspecified — is illusory. There is no univocal thesis of externalism or internalism" (Gertler 2012: 52). In **§4.3**, I will show that such pessimism is unwarranted. Once the distinction between the definition of a debate and substantial accounts within the debate is applied, we can identify a substantive, defining commitment of state internalism and externalism in epistemic terms. In a nutshell, I will argue that what is at stake between state internalists and externalists is whether an individual's mental states may change (or "switch") in a way that the individual is unable to

register entirely out of a certain kind of empirical ignorance and not owing to any of the individual's limitations. We will see that there are reasons specific to the state internalist/externalist debate that invite one to mistake substantial state internalist accounts for the definition of state internalism. This mistake has resulted in the discussion reaching a point where some accounts are not apt to play the theoretical role they were designed to play, and some are dismissed in virtue of misplaced criteria. In **§4.4**, I will conclude this chapter and this thesis by suggesting a unified, epistemic definition of the internalism/externalism debates in the philosophy of mind and epistemology in terms of an individual's capacity for ready discrimination.

4.2 Existing Definitions and Their Issues

4.2.1 The spatiophysical definition

The labels of "internalism" and "externalism" suggest a spatiophysical reading of the respective positions: state internalists about mental states consider purely internal mental states to supervene on states "inside" the thinker whereas state externalists allow for factors "outside" the thinker to contribute to the individuation of mental states. Common to the definitions discussed in this section, the physical definition (**§ 4.2.1.1**) and the spatial definition (**§ 4.2.1.2**), is that they take the boundary between what is internal and external to a thinker relevant to the state internalist/externalist debate to be some physical boundary such as the body or the skin of the individual.

4.2.1.1 The physical definition

The physical definition characterises state internalism positively by reference to the space "within" an individual and state externalism as the negation of state internalism:

> *Physical definition of state internalism*: a theory is state internalist if and only if, according to it, an individual's purely internal mental states at t supervene on the individual's physical states at t.

It follows from the physical definition that, according to state internalists, if two individuals are in the same physical, and particularly in the same brain states, they must be in the same purely internal mental states. State externalism is defined as the negation of state internalism: two individuals may be in the same physical states and yet be in distinct mental states. The physical definition is how the debate has been defined by most theorists (e.g., Putnam 1981; Burge 1988; Jackson and Pettit 1996; Boghossian 1997; Davies 1998; McLaughlin and Tye 1998).

One issue with the physical definition often noted is that it does not allow state internalists to say that, necessarily, Oscar and Twin Oscar are in the same purely internal mental state in the classical Twin Earth scenario: according to Putnam's set-up, Oscar and Twin Oscar are not in the same brain-states for Oscar's brain contains water and Twin Oscar's brain contains twin water. However, despite this difference on the neurochemical level, it is likely that state internalists will take Oscar's water-thoughts and Twin Oscar's twin water-thoughts to satisfy the same functional role.[61] Importantly, the same functional role can be satisfied by different physical substrates in different organisms. So, in order to abstract away from irrelevant physical differences between Oscar and Twin Oscar, state internalists may prefer the following physical definition:

> *Functional definition of state internalism*: a theory is state internalist if and only if, according to it, an individual's purely internal mental states at t supervene on the individual's functional states at t.

---

[61] The individuation of the inputs and outputs will be crucial for the individuation of the functional role itself. As mentioned in chapter 2, the individuation of functional roles is a point of debate between state internalists and externalists. I will put this point aside for this section. The problems that the spatiophysical definition faces would come up even if functional roles could only be individuated internally.

One immediate worry with the functional definition is that it makes (non-reductive) physicalism that accepts the supervenience of the mental on the physical an analytic consequence of state internalism (see, e.g., Block 1997; Fodor 1997). But, as has been pointed out by a number of theorists (Farkas 2003: 189; Gertler 2012: 55), there are certainly dualist state internalist views. Dualists believe that an individual's mental states and any physically realised property, including functional states, either belong to essentially different kinds or instantiate essentially different properties (see, e.g., Descartes ([1641]/1998; Chalmers 1996; Lowe 2006).[62] A dualist state internalist may, for example, claim that an individual is in the same purely internal mental state in and out of a twin case in virtue of the fact that they instantiate the same dualist property. But state internalist dualism is classified as a state externalist view according to the functional definition: any dualist property meets the criterion for the external that underpins the functional definition as that which is not within the body of the individual. On state internalist dualist views, two individuals may be in the same functional states but in distinct mental states because no law-like relation holds between the states of an organism and the organism's mental states according to dualists. So functional sameness is not sufficient for mental sameness on all state internalist views.

Materialistically-minded state internalists and state externalists may not be particularly worried about the fact that the functional definition of state internalism does not capture certain dualist views. But there are other considerations that suggest that the functional definition has misidentified the substantial disagreement between state internalists and externalists. As we will see, we can design a twin case that suggests that *something other* than functional sameness may be sufficient for mental sameness according to state internalism.

Some psychological kinds are functionally individuated. Let us suppose that depression on Earth is a completely psychological disorder which

---

[62] See Parrott and Gomes (2019: n. 13), however, who highlight reasons why Descartes' First Meditation is arguably better understood as a *locus classicus* for a state internalist thought experiment rather than an advancing of the view.

causes an individual to experience negative mood, as well as to be anxious and irritated.[63] Let us further suppose that the cause for depression on Earth is anxiety. On Twin Earth, there exists a very similar psychological disorder that Twin Earthlings call "depression", which also causes individuals to experience negative mood, as well as to be anxious and irritated. On Twin Earth, the cause for what they call "depression" is irritation. It is the year 1750: research on psychological disorders is not developed. Let us suppose that Oscar experiences depression on Earth and Twin Oscar experiences what he calls "depression" on Twin Earth. Suppose as well that Oscar and Twin Oscar are exactly alike in all other respects: they have exactly the same physical microstructure and everything seems the same to them since they are experiencing the same psychological symptoms. But insofar as disorders are individuated by their causes, when Oscar believes "Depression is a monster" and when Twin Oscar believes what he would express in the same way, they will express different propositions according to the state externalist.[64]

The functional definition does not allow us to express state internalism as opposed to externalism about such a twin case. In virtue of the fact that Oscar's disorder is caused by anxiety and that Twin Oscar's disorder is caused by irritation, they are not in the same functional states. But if they are not in the same functional states, according to the functional definition, state internalism does not entail that they are necessarily in the same purely internal mental state when they express the above belief. But it seems that state internalism *should* entail that they are in the same purely internal mental state. It seems that state internalists should claim that "depression" expresses the same concept in Oscar's and Twin Oscar's mouths (and minds) before they or anyone in their community knows about the difference between the two psychological disorders. On the functional definition of state internalism, the case poses no challenge to state internalists but, intuitively, it should.

---

[63] Thanks to Maya Krishnan for the example.

[64] Farkas discusses a similar twin case as an objection to the physical definition (see her 2003: 190; 2008: 76). Since certain objections that are effective against the physical definition are not effective against the functional definition, the twin case I present focusses on the functional definition. Note also that Farkas presents her case as showing that physical sameness is not *necessary* for mental sameness according to state internalism (2003: 190). But necessity was never the issue: it is compatible with the physical/functional definition of state internalism that two individuals may be in the same mental state but are in distinct physical/functional states.

4.2.1.2 The spatial definition

The functional definition is a weak variant of the physical definition that identifies a specific physically-realised property located "inside" the individual, on which state internalists would consider purely internal mental states to supervene. The spatial definition corresponds to a different approach. It characterises state externalism positively with reference to the space *around* the individual, and state internalism as the negation of state externalism:

> *Spatial definition of state externalism*: a theory is state externalist if and only if, according to it, there is a case in which some of an individual's mental states partly supervene on factors located outside of that individual.[65]

State internalism is defined negatively as the position according to which there are no cases in which an individual's primary mental states supervene on factors that are located outside of the individual.

An advantage of the spatial definition is that certain dualist state internalist views come out as state internalist. Take Cartesian dualist state internalism, according to which individuals are in the same purely internal mental state in and out of twin cases in virtue of the fact that water and twin water seem the same to them while facts about how things seem to individuals supervene on an immaterial mental substance (see Descartes [1641]/1998). Such a view will come out as state internalist according to the spatial definition in virtue of the fact that the immaterial substance is not located outside of the individual.

Considering metaphysically extravagant views, however, also reveals the shortcomings of the spatial definition. We can construct another twin case that is such that the spatial definition does not allow us to capture the sense in which state externalism poses a challenge to state internalism.

---

[65] By 'X partly supervenes on Y', I mean 'Y is part of a *minimal* supervenience base for X'.

Consider a Christian community on Earth in which wisdom is constituted by the Holy Spirit's presence to the individual and a Christian community on Twin Earth in which wisdom is constituted by The Virgin Mary's presence to the individual. Neither community knows about the fundamental constitution of wisdom. Let us suppose that Oscar experiences wisdom on Earth and Twin Oscar experiences what he calls "wisdom" on Twin Earth. Suppose as well that Oscar and Twin Oscar are exactly alike in all physical and phenomenal respects. Both believe what they would express with "Wisdom is a virtue". Intuitively, state externalists should say that the thought that Oscar and Twin Oscar thereby express are different because they are individuated with respect two different spiritual entities while state internalists should say that they are in the same purely internal mental state. However, it is not the case that the individual's mental states depend on factors *located* outside of them as neither the Holy Spirit nor the Virgin Mary are located at all. According to the spatial definition, the state externalist view portrayed in this case would be classified as state internalist, which, intuitively, it should not. Relatedly, it would seem that if state internalist dualist views come out as state internalist according to the spatial definition, they do so for the wrong reasons.

The internal/external boundary relevant to the state internalist/ externalist debate cuts cross the spatiophysical internal/external boundary: states of an immaterial soul, essentially distinct from the body, would seem to be relevantly "internal" according to some state internalist views; brain states and functional states would seem to be relevantly "external" according to some state externalist views. This observation comes back to the first noise in Putnam's original Twin Earth case where Oscar and Twin Oscar are not physically identical given the molecular difference between water and twin water in their bodies. This complication is usually brushed off with the observation that Putnam could have used other natural kinds that do not occur in the human body as an example. But that seems beside the point. It is easy to imagine cases of natural kinds that *only* occur within the human body and that form the basis for twin cases that we would expect to divide state internalists and externalists just as Putnam's original case did. To the extent that the

physical and the spatial definition carve up the taxonomy of state internalist and externalist positions in the wrong way, they have misidentified the issue between state internalists and externalists.

4.2.2 The phenomenal definition

The anxious depression/irritable depression and The Virgin Mary/Holy-Spirit twin cases may provide a clue on what is more relevant to the state internalist/ externalist debate than a spatiophysical boundary: the fact that an individual's mental states are not readily discriminable to that individual (assuming the appropriate successive or descriptive modes of presentation). For note that the individual in each twin case is unable to readily discriminate anxious depression-states from irritable depression-states, or states about 'holy wisdom' from states about 'immaculate wisdom'. This may be relevant to the state internalist/externalist debate insofar as state internalists believe that if the individual is unable to readily discriminate between the states, then they are in the same state, whereas the states may be distinct nevertheless according to state externalists.

One prima facie plausible way to achieve a set-up in which some of an individual's mental states are readily indiscriminable to them while the states themselves may vary is to hold the individuals' physical states fixed and vary the environmental facts but it assumes that the right psycho-physical laws hold. A more direct way to ensure that some relevant mental states are readily indiscriminable to an individual while the relevant mental states may vary is to hold certain appearance-states fixed and vary the environmental facts. Common to the definitions that will be discussed in this section, the phenomenal definition (**§ 4.2.2.1**) and the special access definition (**§ 4.2.2.2**), is that they consider the boundary between what is internal and external to a thinker relevant to the state internalist/externalist debate to be the boundary around those states that fix how things appear to the individual.

4.2.2.1 The phenomenal definition

According to many theorists, it is an individual's phenomenal states that determine how things appear to the individual (see, e.g., Block 1995; Chalmers 1996; Farkas 2008). According to such views, if the appearance of a mental state is fixed, then the phenomenal states of that mental state are fixed: "when I say that things appear the same (colour, shape, or otherwise), this amounts to saying that the experiences of things looking in this way for some subjects have a common phenomenal property" (Farkas 2008: 89). On this view, if the relevant mental states are readily indiscriminable to an individual in and out of a twin case, this is because the relevant mental states share certain appearance properties, i.e. they instantiate the same phenomenal states. This forms the basis for the phenomenal definition of the state internalist/ externalist debate:

> *Phenomenal definition of state internalism*: a theory is state internalist if and only if, according to it, an individual's purely internal mental states at t supervene on the individual's phenomenal states at t.

It follows from the phenomenal definition that, according to state internalists, if two individuals are in the same phenomenal states, they must be in the same purely internal mental state. State externalism is defined as the negation of state internalism: two individuals may be in the same phenomenal states and yet be in distinct mental states.

The phenomenal definition of the state internalist/externalist debate has been defended most explicitly and forcefully by Katalin Farkas (2003, 2006, 2008b), and, albeit more indirectly, by Siewert (1998), Horgan and Tienson (2002), and Loar (2003). Most of the latter theorists defend a phenomenal definition of the internalist/externalist debate on thought contents. So-called strong phenomenal intentionalists have consecutively extended phenomenal accounts to all mental phenomena, including attitudes towards contents (see, e.g., Pitt 2004, Farkas 2008a, and Mendelovici 2018), and made a phenomenal definition of the internalist/externalist debate on mental states possible.

An advantage of the phenomenal definition is that certain state internalist dualist views come out as state internalist. Phenomenal states are non-physical, subject-dependent states. As briefly mentioned, some state internalist forms of property dualism embrace phenomenal states as the supervenience base of purely internal mental states, and claim that an individual in and out of a twin case is in the same mental states in virtue of being in the same phenomenal states (see Chalmers 1996; 2002). These views are properly classified as state internalist according to the phenomenal definition.

Another advantage is that, assuming the phenomenal definition, state internalism and state externalism come out as opposing views over the Virgin Mary/Holy Spirit twin case as well as the anxious depression/irritable depression twin case. The difference in the fundamental constitution of wisdom and of depression in Oscar's and Twin Oscar's world is not phenomenally accessible.

But the phenomenal definition has odd implications for certain externalist views on phenomenal states. Recall, phenomenal externalism that was introduced in chapter 1. According to phenomenal externalists, phenomenal states are either identical to, or supervene on, certain externally-individuated representational properties of mental states. The phenomenal definition has two odd implications for phenomenal externalism that are the two sides of the same coin.

First, phenomenal externalism does not satisfy the phenomenal definition of state externalism: it is not the case that, assuming phenomenal externalism, two individuals may be in the same total phenomenal states and yet be in different mental states. To see why, assume phenomenal externalism about perceptual experiences. Since the phenomenal states of a perceptual experience are identical to, or supervene on, the experience's representational properties, perceptual experiences with the same total phenomenal states are necessarily states with the same representational properties, i.e. representations of the same object and its qualities. Thus, any case in which individuals are in the same relevant phenomenal states is trivially a case in

which individuals are in the same perceptual experience: the state of perceiving the same object and its qualities.

Second, by that same fact, phenomenal externalism satisfies what is a direct implication of the phenomenal definition of state *in*ternalism, namely that if two individuals are in the same relevant phenomenal states then they must be in the same mental states. But it is odd that what would seem like a paradigmatically externalist view about some mental condition would satisfy a direct implication of state internalism. Finally, it is a consequence of this oddity that classic twin cases will not divide state internalists and externalists given the phenomenal definition assuming phenomenal externalism: an individual is not in the same total phenomenal states in and out of a twin case according to phenomenal externalists; *a fortiori*, the state externalist challenge cannot consist in claiming that an individual may be in different mental states despite being in the same phenomenal states.

In response to these difficulties, some state internalists may be willing to grant to phenomenal externalists that in some cases in which an individual's mental states are readily indiscriminable to them, they may nevertheless be in distinct phenomenal states. But it is a natural thought that if it is not in virtue of the individual's phenomenal states that their mental states are readily indiscriminable to them, it must be in virtue of some other state that fixes how things appear to the individual. A natural candidate is the states to which an individual has special access. This forms the basis for the special access definition of the debate.

4.2.2.2 The special access definition

An individual has special access to a mental state if and only if the individual has special access to the content as well as to the attitude they take towards that content. Some state internalists may suggest that an individual's mental states are readily indiscriminable to that individual in a context in virtue of the fact that the individual is in the *same specially accessible states* in that context. Like the phenomenal definition, such a view would explain the fact that an individual's mental states are readily indiscriminable to that individual

in and out of a twin case by virtue of the identity of some state of the individual. In virtue of the fact that phenomenal states and specially accessible states are distinct kinds of states, an individual may be in distinct phenomenal states, while being in the same specially accessible states. This allows to formulate a definition of the debate by reference to how things appear to the agent while granting that phenomenal externalism may be correct:

> *Special access definition of state internalism*: a theory is state internalist if and only if, according to it, an individual's purely internal mental states at t supervene on the individual's specially accessible states at t.

The problem with the special access definition is that it attracts the same kind of issues as the phenomenal definition: it has odd implications for certain externalist views on self-knowledge.

I discussed certain externalist theories of self-knowledge in chapter 1 and 2 (§1.4.1; §2.2.1). According to compatibilists, individuals are not in the same specially accessible states in and out of twin cases: Oscar, for example, has special access to his belief that water is transparent whereas Twin Oscar has special access to his belief that twin water is transparent. The special access definition of state internalism has odd implications for such externalist theories of self-knowledge: they do not satisfy the proposed definition of state externalism, but they do satisfy a direct implication of state internalism as defined. A consequence of this oddity is, again, that classic twin cases will not divide state internalists and externalists assuming the special access definition.

Take Alex Byrne's transparency account of self-knowledge (2018). On Byrne's account, the second-order belief about one's states is "transparent" to one's first-order state.[66] Thus, any case in which individuals are in the same specially accessible states is trivially a case in which individuals are in the

[66] Byrne's account systematically develops a view on self-knowledge notably propounded by Evans. Famously, Evans asked the following question about how we self-ascribe beliefs: [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?," I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (Evans 1982: 225)

same mental state. But individuals will not be in the same specially accessible states in and out of twin cases. So the state externalist challenge cannot consist in claiming that individuals may be in different mental states despite being in the same specially accessible states.

The reason why the phenomenal and the special access definitions run into trouble is one we have encountered before. In chapter 2, I discussed whether, while assuming state externalism, one could identify any mental state or property that is uncontroversially shared between an individual in and out of twin cases (and an individual in and out of really bad cases). It turned out that for any candidate state or property, some state externalists will reject the view that the property stays constant across distinct environments.

In the present context, this suggests that it is a bad idea to define the debate between state internalists and externalists as a disagreement over whether mental states supervene on some metaphysically robust mental property. It is a bad idea because it results in false claims and misclassifications: it is not the case that on any state externalist view, individuals may be in distinct mental states while being in the same phenomenal or specially accessible states. With the physical and spatial definitions, we saw that the internal/external boundary relevant to the state internalist/externalist debate cuts cross the spatiophysical internal/external boundary. With the phenomenal and the special access definitions, we find something analogous. Take the phenomenal definition. Assuming the phenomenal definition, the debate between state internalists and externalists collapses into the debate over the relation of the intentional and the phenomenal: the state internalist view is that the intentional and the phenomenal are either identical or very closely related, which is denied by state externalists. But certain state externalist views do not deny a close connection of the intentional and the phenomenal, to the contrary, they consider the phenomenal to be either identical or very closely related to the intentional. Yet, they come to very different verdicts about twin cases. To the extent that the phenomenal and the special access definitions carve up the taxonomy of state internalist and externalist positions in the wrong way, they

too have misidentified the issue between state internalists and state externalists.

4.2.3 The epistemic definition

State internalists and externalists agree that an individual's relevant mental states are readily indiscriminable to the individual in and out of twin cases. They further agree to disagree about whether individuals are in the same mental states in and out of twin cases. What is common to the definitions discussed so far is that they account for the fact that an individual's relevant mental states are readily indiscriminable in and out of twin cases in terms of some robust identity, namely physical, functional, phenomenal, or introspective identity. But we have seen that the metaphysics of the mind are, maybe paradoxically, too contested between state internalists and externalists to define their debate. This leaves the definition that makes the epistemic datum its centre-piece. Common to the epistemic definitions that will be discussed in the subsequent sections is that they take the boundary between what is internal and what is external to a thinker relevant to the state internalist/externalist debate to be the boundary around those facts that are epistemically accessible to the thinker in a certain way.

In recent years, a number of theorists have had positive things to say about the proposition that the state internalist/externalist debate is, at base, a debate about the epistemology of the mind (Farkas 2003; Gertler 2012; Parrott and Gomes 2021). As we will see, an epistemic definition has a number of advantages over the metaphysical definitions (**§ 4.2.3.1**). However, the epistemic definition faces serious issues insofar as state internalism would look to be a hopelessly implausible view from the get-go if it is defined epistemically (**§ 4.2.3.2**). Farkas (2008: Chap.5) and Gertler (2012: 52) do not think that these issues for the epistemic definition can be solved. Parrott and Gomes propose a version of the epistemic definition—the impersonal epistemic definition—in the hope of solving them. But I will argue that, while

Parrott and Gomes' proposal is promising in some respects, it also creates further issues (**§ 4.2.3.3**).

4.2.3.1 The epistemic definition

Relying on the appropriate modes of presentation defended in the previous chapter, I propose the following epistemic definition of state externalism:

> *Epistemic definition of state externalism*: a theory is state externalist if and only if, according to it, there is a case α that is such that S is unable to readily discriminate the case α in which S is in M_α presented to S as "my current state of thinking that *p*" from an alternative case β in which S is in M_β presented to S either as "my earlier state of thinking that '*p*'" or as "my counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which S does not readily know that it is not S's world), and $M_\alpha \neq M_\beta$.[67]

State internalism is defined as the negation of state externalism: there are no cases such that an individual is unable to readily discriminate between their mental states and they are in distinct mental states. So it follows that, according to state internalists, if an individual is in readily indiscriminable mental states, they are in the same purely internal mental state. If we focus on cases in which a twin case is a relevant alternative, i.e. a switching case, the substantial disagreement between state internalists and externalists concerns whether a mental state change may (eventually) take place that is readily indiscriminable to the individual to whom it occurs.

Note, first, that the epistemic definition has a number of taxonomical virtues. State internalist dualist views come out as state internalist in virtue of the fact that they deny that a readily indiscriminable mental state change could

---

[67] I will not mention the modes of presentation explicitly from here onwards. Whenever I will speak of mental states being readily indiscriminable without explicitly mentioning the relevant modes of presentation, I will assume the successive or descriptive modes of presentation discussed in detail in chapter 3.

take place in switching cases, regardless of whether they consider an individual's mental states to supervene on dualist phenomenal properties or an immaterial substance. In fact, historically, it was often the same epistemic assumptions about the mind that ruled out the possibility of readily indiscriminable mental state changes according to state internalists that motivated dualism itself: it is at least partly in virtue of the fact that Descartes considers the mind to be essentially transparent to itself that he rejects a necessary connection between the body and the mind (see Descartes [1641]/ 1998: Second Meditation). Further, state internalists and externalists come out as divided over the anxious depression/irritable depression twin case as well as the Virgin Mary/Holy Spirit twin case as we would expect them to. Any view according to which an individual is in distinct yet readily indiscriminable depression-states, or distinct yet readily indiscriminable wisdom-states in and out of the respective twin case will be classified as state externalist. Whether the individual is in the same physical, functional, phenomenal or specially accessible states in and out of the twin case is incidental.

Note further that the epistemic definition nicely matches findings about twin cases made throughout the chapters, including this one. The epistemic definition of the state internalist/externalist debate corresponds to the view that state internalists and externalists' disagreement over twin cases is, essentially, a disagreement over whether an individual must be in the same mental state in two cases if they cannot activate knowledge of distinctness of the relevant states without significantly changing their epistemic situation. That matches how twin cases are set up: the basis from which any twin case (and any bad case mind you) is developed is to imagine a case that is readily indiscriminable from the actual case because it differs from the actual case only with respect to some hidden aspect. For example, Jessica Brown (2004: 38) says that the counterfactual twin scenario "is set up in such a way that things would seem subjectively just the same to the subject if she were in that environment". Anthony Brueckner (1990: 449) says that, "if I were on Twin-Earth thinking that some twater is dripping, things would seem exactly as they now seem (and have seemed)". Simon Blackburn (1984: 324), describes a

series of Twin Earth style scenarios where “everything is the same from the subject’s point of view”.

Some of these quotes read like something a proponent of the phenomenal definition would like to stress. It is a feature common to the phenomenal and the epistemic definitions that they emphasise the epistemic datum that the relevant mental states are readily indiscriminable to an individual in and out of a twin case. But, as Parrott and Gomes point out (2021: 325), in virtue of the fact that the epistemic definition is neutral with respect to the metaphysical underpinnings of the epistemic datum, it is wider in scope than any of the metaphysical definitions. The epistemic definition allows for cases where an individual’s mental states are readily indiscriminable to the individual but where the individual is nevertheless in different physical, functional, phenomenal or specially accessible states. To read the quotes as supporting a phenomenal conception of the debate, and not just as emphasising the epistemic datum, seems to me to put a stronger claim into these people’s mouths than they may be willing to make.

In chapter 3, I showed that for all switching cases used to illustrate any variety of state externalism, it holds that individuals are unable to readily discriminate a particular occurrent mental state from a relevant alternative mental state. This would also hold for the switching variant of the Virgin Mary/Holy Spirit and anxious depression/irritable depression twin cases: individuals would be unable to readily discriminate an occurrent state of holy wisdom from an alternative state of immaculate wisdom, an occurrent state of anxious depression from an alternative state of irritable depression. The fact that the relevant mental states are readily indiscriminable to an individual in and out of a twin case looks to be the central, non-negotiable element of any twin case.

4.2.3.2 The problem of inability

According to the epistemic definition, state internalism is defined as follows:

> *Epistemic definition of state internalism*: a theory is state internalist if and only if, according to it, if S is unable to readily discriminate the case α in which S is in (the purely internal) M_α presented to S as "my current state of thinking that *p*" from an alternative case β in which S is in (the purely internal) M_β presented to S either as "my earlier state of thinking that '*p*'" or as "my counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which S does not readily know that it is not S's world) at t, then $M_\alpha = M_\beta$ at t.

State internalism such defined licences the inference from an individual's inability to readily discriminate between some relevant purely internal mental states to the identity of those purely internal mental states. Parrott and Gomes say, for example, that the fact that "your twin case is, in some sense, indiscriminable from yours (…) suggests that we can use the notion of indiscriminability to fix the notion of internal sameness" (2021: 325). But it is this inference that would seem to make state internalism as defined by the epistemic definition an impossible view to hold.

The following tolerance principle is a consequence of state internalism as defined above:

Mental Tolerance: for any cases α and β, if an individual's relevant mental states in α are readily indiscriminable to the individual (given the appropriate modes of presentation) from the individual's relevant mental states in β, then the individual is in the same purely internal mental state in α and β.

Williamson has presented sorites-style arguments against similar tolerance principles at different places in his work (2000: Chap. 4; 2007). The same type of argument can be used in order to show that Mental Tolerance is false, or rather, that it is false for thinkers like you and me.

Recall the thought experiment discussed at the end of chapter 2. An individual is presented with a series of coloured patches that incrementally change colour from purple to yellow over some period of time. Each situation

in which the individual is looking at a coloured patch corresponds to a case. In each case, the individual is in a particular perceptual state with the content that the coloured patch is of THAT colour at that moment (they refer to the colour demonstratively). The colour of the patch in each case is readily indiscriminable to the individual from the colour of the immediately succeeding patch. Consequently, the individual's perceptual state in each case is readily indiscriminable to that individual from their perceptual state in the immediately succeeding case given a successive mode of presentation: the individual is unable to readily discriminate their perceptual state at t presented to them as "my state of seeing that *p* just a moment ago" from their perceptual state at t_{+1} presented to them as "my current state of seeing that *p*".

According to Mental Tolerance, for any two cases α and β, if the individual's mental states in α are readily indiscriminable to that individual from their mental states in β (given a successive mode of presentation), then the individual is in the same mental state in α and β. According to Mental Tolerance, the individual is in the same mental state at t_{+1} as they were at t; in virtue of the fact that the mental state they are in at t_{+1} is readily indiscriminable to them from their mental state at t_{+2} (given a successive mode of presentation), they are in the same mental state at t_{+2} as at t_{+1}; and so on. It follows from Mental Tolerance that the individual is in the same mental state in the last case as they were in the first case. But they are not in the same mental state in the last case and in the first case: in the last case, the individual is in the perceptual state of seeing a yellow patch whereas they were in the perceptual state of seeing a purple patch in the first case. Therefore Mental Tolerance is false.

The systematic issue behind the failure of Mental Tolerance is the following. Mental identity, like identity about any property, is transitive whereas ready indiscriminability, like any form of indiscriminability, is non-transitive. For cognisers with limited discriminative abilities like you and me, indiscriminability is not a reliable criterion for identity in all contexts because in certain contexts, distinct entities will be indiscriminable to us.

The sorites series of readily indiscriminable mental states highlights a more general issue with a state internalist account of mental identity in terms

of an individual's inability to readily discriminate between relevant states. There are a number of reasons for which an individual may be unable to readily discriminate an occurrent mental state from a relevant alternative state that are extraneous to the debate on mental state individuation. Individuals may be drunk, tired, or distracted which may temporarily cause an occurrent mental state to be readily indiscriminable to them from a relevant alternative state that would not normally be readily indiscriminable to them from the relevant alternative state. Recall from chapter 2: if we ask the person presented with the sorites series to discriminate between the colours under stress, some colours are likely to be readily indiscriminable to them that would not normally be readily indiscriminable to them. The epistemic definition of state internalism implausibly predicts in such a case that the two readily indiscriminable states are the same purely internal mental state when the individual is distracted but are distinct when they are not. A state internalist view according to which it is impossible for individuals to lack discriminatory ready access to their distinct mental states in any circumstances is extremely implausible.

We can control for such temporary disruptions to an individual's ability to readily discriminate by appealing to the difference between general and specific abilities. Recall from chapter 3 (§3.4.2), general abilities, in contrast with specific abilities, are those abilities that individuals have stably, across a number of contexts. Importantly for the present context, unfavourable temporary circumstances do not rob people of their general abilities. My general ability to do a handstand, for example, may be masked by the fact that my arm is broken. But, here and now with my broken arm, I still had the relevant training and have done many handstands in the past when I tried to do it. In a similar way, an individual's general ability to readily discriminate between their current mental state and a relevant alternative mental state may be masked by their distraction, tiredness or drunkenness. So maybe state internalists should not rely on an individual's *specific* inability to readily discriminate between some relevant states as a criterion for the identity of those states but the individual's *general* inability to readily discriminate between the relevant states:

> *General epistemic definition of state internalism*: a theory is state internalist if and only if, according to it, if S is *generally* unable to readily discriminate the case α in which S is in (the purely internal) M_α presented to S as "my current state of thinking that *p*" from an alternative case β in which S is in (the purely internal) M_β presented to S either as "my earlier state of thinking that '*p*'" or as "my counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which S does *generally* not readily know that it is not S's world) at t, then $M_\alpha = M_\beta$ at t.

The issue with the general epistemic definition is that it is also subject to counterexamples of a similar kind. In some contexts, an individual is generally unable to readily discriminate between their occurrent mental state and a relevant alternative mental state, and yet, even to the mind of state internalists, it would be false to conclude that they are in the same purely internal mental state. Consider the sorites series of coloured patches just discussed. A human individual will be generally unable to readily discriminate between the patches of two successive colours if the similarities fall below a certain threshold of what the human eye can discern. The same applies to the corresponding perceptual states. Yet, the only way to account for the fact that the individual is in distinct states at the beginning and the end of the series is if, throughout the series, the individual is in distinct, yet readily indiscriminable perceptual states. We encountered other examples of the same kind before. Recall the person in the curtain shop who cannot readily discriminate the colour of the curtains in front of them from the colour of their carpet at home because they cannot remember the colour of their carpet (which we may suppose is distinct from the colour of the curtains) with the sufficient level of detail. There is a natural limit to the phenomenal detail (cognitively normal) human adults can retain in long-term memory, and it is likely to be less than the level of phenomenal detail than they are able to perceptually discriminate. Again, the general epistemic definition of state

internalism implausibly predicts that in such cases the two readily indiscriminable perceptual states are the same mental state when they clearly are not.

An individual's general ability to readily discriminate between states is constrained by further cognitive abilities, such as their general ability for colour discrimination or memory. For people like you and me, these abilities are systematically limited in a way that, in some contexts, causes distinct mental states to be readily indiscriminable to us. So not even our general inability to readily discriminate between some relevant states is a reliable criterion for their identity. Let us call this the "*problem of inability*" for the epistemic definition of state internalism. There are many reasons for which individuals may be unable to readily discriminate between some relevant states, some temporary and some permanent, but in many cases extraneous to questions of mental state individuation.[68] It is unclear how to pick out the *right kind of inability* to readily discriminate between one's states, if there is one, that could plausibly serve as a criterion for the identity of the relevant states according to state internalists.

Note that state internalism does not get into these issues when it is defined along the lines of the metaphysical definitions. State internalism, metaphysically defined, proposes to use the *identity* of an individual's physical states or some non-intentional mental property as the criterion for the *identity* of that individual's purely internal mental states. State internalism, metaphysically defined, is not committed to false claims about the sorites series of mental states, for example. An individual is arguably not in the same physical or phenomenal states from one case to the next in such a series, even if they may mistakenly believe that they are. One of Farkas' main arguments for why the state internalist/externalist debate must be defined in phenomenal terms is that she does not see how proponents of an epistemic definition could solve the problem of inability (Farkas 2008: Chap.5).

---

[68] Some have pointed out that state internalism as defined by the epistemic definition is hopelessly over-intellectualised, and actually speciesist and ableist, as many creatures are generally unable to readily discriminate between their mental states because they are generally unable to *conceive* of their mental states (see Siegel 2004; Farkas 2008: Chap.5). The objection would be pertinent if state internalists were trying to give an account of the identity of dogs' states, say, by reference to dogs' inability to readily discriminate between their mental states. But as I will argue in §4.3, state internalists are not trying to do that even if their view is correctly defined in epistemic terms.

4.2.3.3 The impersonal epistemic definition

Parrott and Gomes (2021) set out to show how it may be done. If we think of Oscar in the original Twin Earth scenario, then it is not because of any temporary cognitive failings or permanent cognitive limitations that Oscar is unable to readily discriminate between his current water-thoughts and his counterfactual twin water-thoughts. By hypothesis, water is not distinguished from twin water except by its molecular structure. So, it would seem that not only Oscar could not readily discriminate between water and twin water, but no one could. Relatedly, it would seem that not only could Oscar not readily discriminate between water-thoughts and twin-water-thoughts, but no one could. Parrott and Gomes believe that this is the key to solving the inability problem of the epistemic definition of state internalism.

Parrott and Gomes propose conceiving of the inability to readily discriminate between one's states relevant to the state internalist/externalist debate not as the inability of any specific individual, like Oscar, but *impersonally*. They follow Martin (2006) in suggesting that the impersonal conception of an inability prescinds from any reference to actual persons, and their individual- and species-related, permanent and temporary features. They illustrate the impersonal conception of an inability by an example from vision:

> "Suppose that we wanted to fabricate a bunch of replica apples. If we had a sufficient level of skill, and designed the right equipment, we might succeed in producing schmapples, perfect replicas that are visually indiscriminable from genuine apples. Suppose further that not only can we not visually distinguish them, but that no possible visual system can discriminate schmapples from apples. They are impersonally visually indiscriminable" (*ibid.*: 333).

According to Parrott & Gomes, the impersonal conception of an inability is equivalent to an unrestricted impossibility claim (*ibid.*:330/1). We may define an impersonal inability as follows:

> An individual S is *impersonally* unable to ϕ if and only if it is impossible (for any possible individual in any possible circumstances) to ϕ.

Parrott and Gomes suggest that the inability problem for the epistemic definition can be solved by relying on the impersonal conception of the inability to *introspectively* (in my terminology "specially") discriminate between one's mental states. As we will see below, Parrott and Gomes' account would be fairly obviously implausible if cast in terms of ready discrimination. However, as we will also see below, it is implausibly strong in their preferred version as well.

Put in the present framework, their proposition is that the state internalist position is better epistemically defined as follows:

> *Impersonal epistemic definition of state internalism*: a theory is state internalist if and only if, according to it, if the case α in which S is in (the purely internal) M_α presented as "S's current state of thinking that '*p*'" is *impersonally introspectively indiscriminable* from an alternative case β in which S is in (the purely internal) M_β presented as "S's earlier state of thinking that '*p*'" or as "S's counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which *no possible system* introspectively knows that it is not S's world) at t, then $M_\alpha = M_\beta$ at t.[69]

According to the impersonal epistemic definition, the state internalist position is that if an individual is in impersonally introspectively indiscriminable mental states (given the appropriate modes of presentation), they are in the same purely internal mental state. The state externalist position

[69] Parrott and Gomes do not define the state internalist/externalist debate directly but try to identify the correct account of "internal sameness" that, according to them, is at stake in the debate between state internalists and externalist. Parrott and Gomes defend the following epistemic account of *internal sameness*: "Therefore, we propose the following epistemic account of internal sameness: a state S is narrow if and only if, for all cases α and β, if the total state of the individual in α is impersonally introspectively indiscriminable from the total state of the individual in β, then the condition that one is in S obtains in α if and only if the condition that one is in S obtains in β" (2021: 331).

would hold in contrast that an individual may be in states that are impersonally introspectively indiscriminable and yet distinct.

The impersonal epistemic definition is not targeted by the objections to the epistemic definitions discussed earlier. Recall the person in the curtain shop who forgot to bring a sample piece of their carpet to compare colours. The ability for introspective discrimination between present and past states, impersonally conceived, is not constrained by a particular individual's limitations on their phenomenal memory. Consequently, the individual's perceptual states are impersonally introspectively discriminable. The impersonal epistemic definition rightly predicts that, according to state internalists, the individual may be in distinct purely internal mental states in the two cases. The same treatment applies to the sorites series of perceptual states: the perceptual states throughout a sorites series are impersonally introspectively discriminable because the ability for introspective discrimination, impersonally conceived, is not constrained by the limitations on the human ability for colour discrimination. The impersonal epistemic definition of state internalism, again, rightly predicts that, according to state internalists, the individual may be in distinct purely internal mental states throughout the series. The same considerations apply to cases where an individual fails to introspectively discriminate between some relevant states because they are tired, distracted or drunk.

Importantly, according to Parrott and Gomes, Oscar's and Twin Oscar's water- and twin water-thoughts are introspectively indiscriminable even according to an impersonal conception of the inability. Oscar's water-thoughts and Twin Oscar's twin water-thoughts are not just too similar to be distinguished by human cognisers; after all, water and twin water are exact lookalikes except for their molecular structure. So Parrot and Gomes conclude that the impersonal epistemic definition rightly predicts that, according to state internalists, Oscar and Twin Oscar are in the same purely internal mental state. Note that, at least *prima facie*, the impersonal conception of the inability to introspectively discriminate between states constitutes a plausible solution to the problem of inability. The impersonal inability to introspectively discriminate between some relevant states would seem like a reliable criterion

for those states' identity precisely because any mistakes due to a particular individual's exercise of the abilities are excluded from interfering.

At a second glance, however, things start to look different. There are a number of issues with the impersonal epistemic definition of state internalism. Interestingly, Parrott and Gomes are not very worried about the most serious one. They are very worried about certain explanatory gaps that the impersonal epistemic definition seems to generate. But, as I will very briefly point out here and show in detail in §4.3.2.2, these are fairly straightforwardly taken care of once we distinguish between the definition of the debate and substantial accounts within the debate. The most serious issue for Parrott and Gomes' account is the impersonal conception of the inability to introspectively discriminate between states itself.

While the supposed fact that Oscar's water-thoughts are impersonally introspectively indiscriminable from Twin Oscar's twin water-thoughts may constitute a good *criterion* for the identity of those states, it would provide a poor *explanation* for why Oscar and Twin Oscar are in the same purely internal mental states. It is not going to be in virtue of the fact that someone with a very different set of abilities to Oscar and Twin Oscar—a very special set of abilities, unconstrained by any individual- or species- related features—is unable to introspectively discriminate between their (own) states in the same circumstances, that Oscar and Twin Oscar, with their very personal set of abilities, are in the same purely internal mental states according to state internalism. No inability (or ability for that matter), *impersonally* conceived, will provide an explanation for some specific *person's* mental states.

This is the first explanatory gap of the impersonal epistemic definition. In §4.3.2.2, I will suggest that substantial state internalist accounts will explain that Oscar and Twin Oscar are in the same purely internal mental states by virtue of some robust identity, such as an identity of their physical, functional, phenomenal or specially accessible states, although different state internalist accounts will consider different states apt for the task.

The impersonal epistemic definition seems not only to struggle to explain why Oscar and Twin Oscar are in the same purely internal mental state according to state internalism. It also lacks an explanation for the impersonal

inability to introspectively discriminate between some relevant states itself. In virtue of what are certain mental states impersonally introspectively indiscriminable? The obvious suggestion is that the relevant mental states are impersonally introspectively indiscriminable in virtue of having certain properties in common. Parrott and Gomes illustrate the thought by an analogy with visual indiscriminability. Recall schmapples, the replicas of apples so perfect that no possible visual system could visually discriminate them from apples. Parrott and Gomes write:

> "The fact that schmapples are impersonally visually indiscriminable from apples looks to be explained by the fact that they share certain basic visible properties with apples. If schmapples had nothing in common with genuine apples, it would be very hard to grasp why no possible visual system could distinguish the two" (2021: 333).

In the current context, taking the original Twin Earth scenario as example, Twin Oscar's twin water-thoughts are the "schmapples" to Oscar's water-thoughts. Accordingly, one would expect an explanation of why Oscar's and Twin Oscar's states are introspectively indiscriminable for any possible perceptual-cognitive system to turn on features of these mental states themselves. For example, if Oscar's water-thoughts had all phenomenal properties in common with Twin Oscar's twin water-thoughts that would explain why they are impersonally introspectively indiscriminable.

But a definition of the state internalist/externalist debate cannot appeal to any robust properties that Oscar's water-thoughts would have in common with Twin Oscar's twin water-thoughts because any such appeal is premised on specific, controversial views on the nature of such properties. Integrating phenomenal properties, for example, into the definition of state internalism, we have seen, results in misclassifications of certain views over relevant twin cases. But then it looks like the impersonal epistemic definition has no explanation to give for why the relevant mental states are impersonally introspectively indiscriminable. This is the second explanatory gap for the impersonal epistemic definition.

Parrott and Gomes consider the second explanatory gap to be a serious problem for the impersonal epistemic definition (333). It would seem that the epistemic account has to choose between providing a substantial explanation

of why the relevant mental states cannot be discriminated by any possible cognitive-perceptual system, which is likely to result in having to give up the epistemic account in favour of a metaphysical definition (with the problems these tend to bring with them); or somehow showing that no such explanation is required. As Parrott and Gomes point out, the second option would be a way of taking the fact that some relevant mental states are impersonally introspectively indiscriminable to be basic (*ibid.*). Although this may seem to be in the spirit of epistemic accounts of mental phenomena, critical voices against mysterious ungrounded epistemic facts have been growing louder even among those attracted to the theories out of which such epistemic accounts were born (see, e.g., Moran 2019).[70]

Even though they dismiss the second option, Parrott and Gomes believe that they can avoid having to settle for some metaphysically robust explanation. I said before that one would expect an explanation for why some relevant mental states are impersonally introspectively indiscriminable to turn on features of those mental states. Parrott and Gomes suggest that this may be the wrong kind of explanation to look for in the present context. Instead, they say that the impersonal conception of an inability may reveal some sort of absolute limitation on the ability itself. They write:

> "Limitations on the capacity for introspective discrimination should not be explained in terms of any features of the objects of introspection, but in terms of the nature of that capacity itself. In other words, it is the capacity of introspective discrimination, impersonally conceived, that is limited in terms of what it is able to discriminate. And an explanation of those limits need only appeal to the character of introspection itself, rather than to the character of the objects over which it ranges" (2021: 334).

So Parrott and Gomes suggest that a solution to the second explanatory gap may be provided by the impersonal conception of the inability to introspectively discriminate between states itself because abilities may be

[70] Naïve realism is often considered to be committed to 'negative' accounts of hallucinations according to which hallucinations are defined in purely epistemic terms as an experience that is distinct but potentially indiscriminable to the individual from a perceptual experience (Martin 2004; Pautz 2011). But even many attracted to naïve realism find such purely 'negative' accounts of hallucinations inadequate (see Sturgeon 2008b; Moran 2019).

limited in ways that are unrelated to the objects to which they are applied. That sounds fairly mysterious (especially without further elaboration or examples of other abilities whose limitations are specified independently of the objects over which they range). As I will show (§4.3.2.2), there is a more straightforward solution to the second explanatory gap once we distinguish between the definition of the debate and substantial accounts.

The problem with Parrott and Gomes' impersonal epistemic definition is not the explanatory gaps. The problem with their account is the impersonal conception of the inability to introspectively discriminate itself. It is doubtful that we will be able to fully make sense of the impersonal inability to introspectively discriminate but, more importantly in the present context, it does not solve the problem of inability: we have no reason to think that Oscar and Twin Oscar's states are impersonally introspectively indiscriminable.

Parrott and Gomes like to draw on analogies with vision because it might seem that we have a good intuitive grip of when two things are impersonally visually indiscriminable. But it is not that clear that we do. What are the absolute limitations on the ability to visually discriminate between entities? Vision, in the most general definition, is the ability of an individual to interpret their surroundings with the help of the light reflected by the objects in the environment. Of course many more objects will be impersonally visually discriminable than are visually discriminable to us: actual visual systems use ultraviolet and infrared light to interpret their environments. Following these considerations, the insides of two black holes are plausibly impersonally visually indiscriminable. But molecules do reflect light. So water and twin water are visually discriminable to some possible visual systems: those with the ability for molecular vision. While it is unlikely that the discriminations of a system with molecular vision will be able to explain how our human visual states are individuated (see the first explanatory gap), it can seem like they would constitute a good *criterion* for the identity of human visual states: if the system with molecular vision is unable to visually discriminate between two things, we definitely are as well. But in the Twin Earth scenario, such a system *is* able to visually discriminate between water and twin water. And if that is the case, we simply do not know whether

someone with our (and Oscar's) visual abilities will be in the same states or not because we will definitely sometimes be in the same visual state when the system with molecular vision is not. To *us*, the impersonal inability to visually discriminate between things turns out to be a poor criterion for the identity of our visual states.

Before I get to the impersonal inability to *introspectively* discriminate between states, note that the system with molecular vision would be able to *readily* discriminate between water-thoughts and twin water-thoughts. This is because the system draws on the perceptual evidence it has in order to readily discriminate between their states. More generally, it will be very hard to show that some entities are impersonally readily indiscriminable. Ready access, we know, draws on all the evidence that is readily available to a system in a context. But some possible system will always have some information about the objects of discrimination. Take schmapples. Maybe schmapples are impersonally *visually* indiscriminable. It is still unlikely that they are impersonally *readily* indiscriminable; some system just fabricated them and draws on that knowledge to discriminate between them. Similarly, it is likely that some mental states that are impersonally *introspectively* indiscriminable are impersonally readily discriminable. Parrott and Gomes are right to focus on the impersonal inability to *introspectively* discriminate between states.

Yet, our intuitive grip on when two mental states are impersonally introspectively indiscriminable is likely to be blurrier than they make out. Humans discriminate introspectively between fairly heterogenous mental phenomena such as beliefs, desires, intentions, emotions, perceptions, sensations and memories. One may sensibly doubt, and some have doubted (see, e.g., Goldman 1993; Nichols and Stich 2003), that humans employ the same introspective faculty to introspectively discriminate between all of these phenomena. It is even less certain that any possible individual would use the same introspective faculty that we use to find out about their minds. But even if they did, it is unclear what the absolute limitations on this universal ability for introspective discrimination would be. Most importantly, it is unclear whether those limitations could be specified independently of the features of the mental states they range over. On certain accounts of self-knowledge, it is

likely that limitations on the ability to introspectively discriminate between mental states will turn on features of the states being discriminated. For example, on Byrne's transparency account of self-knowledge mentioned earlier, the ability to discriminate between one's states will be directly informed by whether one is able to discriminate between the things and features that these states concern.

Parrott and Gomes do not elaborate on any of these questions: How should we conceive of an absolute limitation on an ability? Why can we assume that every ability is so limited? And why can we assume that we will be able to sensibly determine what these limitations are, given that an impersonal inability is not any particular individual's inability? That's not enough to solve the problem of inability.

In sum, the search for a definition of the state internalist/externalist debate results in a dilemma and in a problem.

The dilemma is the following. The definition of the state internalist/ externalist account cannot postulate any metaphysically robust properties on pain of misclassifying views over relevant twin cases and entangling the debate in issues that are orthogonal to it. The epistemic definition is the only definition that avoids these issues, but it results in an implausible definition of state internalism and multiple explanatory gaps. That is the first horn of the dilemma. The most straightforward solution to these issues—viz. giving a plausible state internalist account of mental identity; closing the explanatory gaps—points back to postulating certain robust properties that would explain the relevant states' epistemic features. But such a solution is unavailable because the metaphysical definitions that postulate such properties in the definition of state internalism are inadequate as definitions of the state internalist/externalist debate. That is the second horn of the dilemma.

As I will show, the distinction between an epistemic definition of the debate and substantial metaphysically committing accounts will dissolve the dilemma. If we know that the dilemma is going to be taken care of, we can focus on solving the problem of inability for the epistemic definition of the debate. It is clear that the epistemic definition requires some qualification: state internalism cannot be the simple-minded view that whenever an

individual is unable to readily discriminate between some relevant states, they are in the same state.[71] Parrott and Gomes' impersonal epistemic definition is meant to provide that qualification but it is premised on a questionable conception of inabilities that furthermore does not provide the correct result in twin cases. In the next section, I will put forward a new epistemic definition of the state internalist/externalist debate, the *categorical epistemic definition* that solves the problem of inability. Then, I will show how it is compatible with a number of different substantial accounts that have been offered on each side of the disagreement.

4.3 An Epistemic Definition, Metaphysical Accounts

In the last section, we could see the following dynamic. The metaphysical definitions formulate plausible versions of state internalism but result in misclassifications of positions over relevant twin cases, and implausible predictions. The epistemic definition sorts the views correctly over all relevant twin cases, but struggles with formulating a plausible version of state internalism. None of the definitions manages to do both. The metaphysical definitions say too much: they define the views in substantial and controversial ways. But the epistemic definition does not say enough to provide the basis for a plausible state internalist account of mental identity.

The reason, for this, I believe is that the definitions of the state internalist/externalist debate have been taking on two conflicting jobs at once: they define the state internalist/externalist debate as well as—or rather, by means of—stating minimal substantial accounts. This is particularly conspicuous in those cases in which the debate is defined by way of defining state internalism positively which is then denied by state externalism. Take the

[71] In the same way, hybrid epistemic internalism (recall from §2.4.2) cannot be the simple-minded view that whenever some individual is unable to readily discriminate between facts about their evidence, i.e. certain relevant propositions and bodies of evidence, that they have justification to believe those propositions to the same extent.

phenomenal definition, for example. The debate is defined as concerning the issue whether an individual's mental states supervene on that individual's phenomenal states by way of stating the state internalist view according to which an individual's mental states supervene on the individual's phenomenal states. But the definition of state internalism has to implicitly assume an internalist conception of phenomenal states in order to work as a definition of state internalism. It is no surprise that misclassifications of other views follow as soon as such assumptions are questioned.

Here I will suggest separating the jobs: it is one task to define the state internalist/externalist debate, and it is a different one to provide plausible substantial accounts of positions within the debate. Once we apply the distinction between a definition and substantial accounts, we will be able to assess accounts with respect to the theoretical role they are apt to play, and without applying misplaced criteria.

To that end, I will propose the categorical epistemic definition that satisfies all of the desiderata on a good definition (**§4.3.1**). I will show how the spatiophysical and phenomenal definitions, if read as substantial state internalist accounts, help to prevent any explanatory gaps from opening (**§4.3.2**). I will close by giving a diagnosis of why the distinction was missed in the debate on the definition of the state internalist/externalist debate specifically. We will see that once the combination of the categorical epistemic definition and substantial accounts is applied, a number of things fall nicely into place (**§4.3.3)**.

4.3.1 The categorical epistemic definition

An epistemic definition of the state internalist/externalist debate in terms of an individual's capacity for ready discrimination needs to solve the problem of inability. The problem of inability was the problem of how to pick out the right kind of inability to readily discriminate between some relevant states that could plausibly serve as the criterion for the identity of those states, according to state internalists. Here I will defend the view that it is an individual's

categorical inability to readily discriminate between some relevant mental states that is at stake between state internalists and externalists (**4.3.1.1**). This will form the basis for the categorical epistemic definition that I will put forward. As we will see the categorical epistemic definition identifies a plausible substantial disagreement between state internalists and externalists, sorts views correctly and predicts their implications plausibly (**§4.3.1.2**).

4.3.1.1 Solving the problem of inability

The task is to find a conception of an individual's inability to readily discriminate between their states that hits the following sweet-spot. On the one hand, the required qualification needs to be such that it rules out cases where the individual fails to readily discriminate between their states in virtue of temporary failings or permanent limitations. So there must be some amount of idealisation involved. On the other hand, the idealisation cannot go so far as to result in an inability to readily discriminate between certain relevant states that has not got anything to do with the particular individual since the state internalist/externalist debate is, after all, a disagreement about how the states of particular individuals are individuated.

We saw that we can abstract away from temporary limitations to an individual's ability to readily discriminate between their states by focussing on the individual's general instead of their specific discriminative abilities. The harder part is to capture, on the one hand, the difference between those cases in which individuals are generally unable to readily discriminate between their states, where state internalists and state externalists *agree* that the individual is in different states, and, on the other, those cases in which individuals are also generally unable to readily discriminate between their states but where state internalists and externalists *disagree* about whether or not the individual is in the same state. In order to find that sweet-spot, let us have a look at its focal point: the difference between an individual's inability to readily discriminate between their states throughout a sorites series of mental states and an individual's inability to readily discriminate between their states in and out of a twin case.

Let us focus on the same sorites series of perceptual states discussed earlier: an individual is looking at coloured patches that change colour incrementally from purple to yellow. What is special about a sorites series of perceptual states of this kind is that the individual's inability to readily discriminate between two consecutive perceptual states of the series is due to limitations on their ability to visually discriminate between colours, and relatedly, perceptual states. In the case of humans, the ability for visual discrimination is limited by a certain threshold of similarity. If the similarity between two particulars falls below that threshold, a human perceiver is unable to visually discriminate between the patches and the corresponding perceptual states. However, if we idealised the human ability for visual discrimination, the entire sorites series of colours as well as of the corresponding perceptual states would become readily discriminable to the perceiver, assuming that the entire series is set in the spectrum visible to humans.

What I mean by that is the following. A human perceiver is able to readily discriminate certain states in the series, namely those that are far enough removed from each other across the series. An *ideal human* perceiver, i.e. a perceiver with the general abilities of (normally developed) humans but without the same limitations on these general abilities, is able to readily discriminate between any particular colour and state in the series, as long as all the colours displayed fall within the spectrum visible to humans. When we idealise the human ability for visual discrimination, we hold the input-level fixed, i.e. the information that hits the retina, and idealise things like the individual's processing-power and attention. The performance of the cone cells in the human retina, as well as the visual and cognitive processing may be improved but there is an in-built limit to the kind of performance improvement possible: the cone cells in humans (or mammals) only respond to wavelengths within a certain spectrum. So when we idealise in the way envisaged, we keep the genetic make-up of humans fixed, *except* for the limitations on interpreting the entirety of the information given at the input-level. One could object that removing certain such limitations would change the input-level itself: for example, if I learn to distinguish between more herbs

and flowers, I am likely to see more of them. But even if removing certain limitations changed the input-level to a certain extent, there are in-built limits to the nature of the information that will become visually available to humans in virtue of the structure of the human eye. Those are held fixed because changing those would require more dramatic changes to the human genetic make-up.

In other words, assuming that the specific sorites series is set entirely within the spectrum visible to humans, there is a *gradual* improvement of the general human ability for visual discrimination, namely an improvement such that the ability for visual discrimination matches the actual difference in colour. The improvement would make any colour (and any corresponding perceptual state) of the series readily discriminable to a human perceiver from any other colour (and state) in the series. This also means that *more of the same kind* of visual evidence would make any colour in the series, and relatedly any state in the series, readily discriminable from any other colour and any other state to an ideal human perceiver.

Now compare this to an individual's mental states in and out of a twin case. Let us focus on the switching case of the classical Twin Earth scenario, since the switching case compares more straightforwardly to the change in mental states in a sorites series of mental states. For the sake of the comparison, I will assume state externalism, so we assume that a mental state change takes place in the switching scenario. In chapter 3, I showed that Oscar is unable to readily discriminate between his water-thoughts and his twin water-thoughts given the appropriate modes of presentation. But Oscar's inability to readily discriminate between water and twin water, and the corresponding states, is not owing to the fact that water and twin water, and the corresponding mental states, are too similar to be readily discriminated by Oscar. The problem is not that water and twin water come in water and twin water shades, or in a water and twin water intensity, that are too similar to be readily discriminated by a human perceiver. Note that it could be a problem of some individuals of the species with molecular vision that water and twin water are too similar to be visually discriminable to *them* (it may fall below *their* threshold of molecular discrimination). Water and twin water share all of

their surface properties, so it is not unlikely that the molecular structure of water and twin water is very similar (even though the labelling "XYZ" would misleadingly suggest otherwise). But no human, not even an ideal one, could visually discriminate between different molecules because the wavelengths of the light visible to humans is too long for molecular dimensions. Nor could we improve Oscar's powers of discrimination by taste to the point that he could taste the molecular difference between water and twin water: humans do not have taste buds that discriminate on the molecular level when the surface properties are stable, even if we improved the sensitivity of Oscar's taste buds for the five recognised tastes to the maximum. Nor is the problem that Oscar has some other piece of knowledge that would put him in the position to readily discriminate between water and twin water, and the related thoughts, were he a better reasoner or had better memory. In short, it is not owing to any kind of failing or limitation that Oscar is not able to readily discriminate between his water-thoughts and twin water-thoughts in the switching scenario.

In other words, holding fixed everything about the case, from Oscar's genetic design (minus its limitations on interpreting the input) to his chemical knowledge, no idealisation on Oscar's abilities would put Oscar in a position to readily discriminate between water-thoughts and twin water-thoughts in a switching case. Only a *categorical* improvement of Oscar's epistemic situation, such as *new* chemical evidence about the two substances or a new faculty such as molecular vision or molecular taste, would make water and twin water, and the corresponding states, readily discriminable to Oscar. Put differently, not even an *ideal counterpart* of Oscar, i.e. someone just like Oscar in Oscar's context but without Oscar's limitations to his capacities, would be able to readily discriminate between their water- and twin water-thoughts in a switching case without the relevant piece of new evidence.

I propose the following terminological distinction to capture the difference between an individual's inability to readily discriminate between their states throughout a sorites series of mental states and the individual's inability to readily discriminate between their states in and out of a twin case: an individual is *generally* unable to readily discriminate between their states throughout a sorites series of mental states, but an individual is *categorically*

unable to readily discriminate between their states in and out of a twin case. I will define the distinction as follows:

> An individual S is *generally* unable to ϕ at t if and only if S does not ϕ across a number of relevant circumstances in which S intends to ϕ at t.[72]

> An individual S is *categorically* unable to ϕ at t if and only if S's ideal counterpart does not ϕ across a number of relevant circumstances in which S intends to ϕ at t.

Oscar is generally unable to readily discriminate between his states throughout the sorites series of perceptual states because, whatever the context in which Oscar intends to register the change from one colour to the immediately succeeding colour, as well as the corresponding perceptual states, he will not succeed in registering the change. But an ideal counterpart of Oscar would register the changes because their ability for colour discrimination matches the actual difference in colour in the spectrum visible to humans. But an ideal counterpart of Oscar in the particular context of being chemically ignorant, will not be able to readily discriminate between water-thoughts and twin water-thoughts, across any number of circumstances, unless they receive new evidence. This is because molecular differences between liquids lie beyond what humans, even ideal ones, are able to perceive and discriminate in a context in which they are ignorant of them.

The ideal counterpart of an individual is the result of idealising on the species-and individual-specific general abilities of an individual. In the case of visual discrimination for healthy humans, for example, we hold fixed the visible spectrum for humans and abstract away from the limitations to the ability for visual discrimination in that spectrum. In the case of an individual's general ability to do a handstand, we hold fixed the shape of that person's

---

[72] In §4.2.3.2, it was noted that the difference between an agent's specific and general abilities is that the latter, but not the former, are those abilities that an agent keeps across a number of contexts. Since I am interested in agential abilities, these contexts will be selected from the set of contexts in which an agent *intends* to act in the way that requires the ability in question (see Jaster 2020: Chap.1).

body, gravity etc., and idealise on features such as muscular structure, balance, and potentially other limitations such as the need for food and sleep etc. The ideal counterpart of an individual with the general ability to do a handstand is able to stand on their hands as long as they intend to stand on their hands.

An individual's categorical *in*abilities are those general inabilities that persist through the idealisation on the individual's general abilities in a context: if one abstracts away from any limitations to the individual's general abilities in a context, the individual is still generally unable to do certain things in that context. Those are the things the individual is *categorically* unable to do. To take the clearest example: in the case of human cognisers, no amount of idealisation on the general abilities of any specific human cogniser will turn them from a non-omniscient cogniser into an omniscient cogniser. With respect to sense-perception, humans come with certain in-built limits on what kind of inputs could become available to them. With respect to central cognition, there are presumably limits on the processing-power set by the organic hardware. The ideal counterpart to any human cogniser draws on the (ideal) general human faculties to learn that which they are in a position to know readily in a context; but they will be generally unable to know certain empirical truths without further empirical evidence.

The attribution of a categorical inability to an individual is relative to the set of general abilities of that individual at a time. What an individual is categorically unable to do may differ between the same individual in different contexts, as well as between individuals of the same species, and as well as between species. In a context in which Oscar is not chemically ignorant in a switching case, for example, his water-thoughts and twin water-thoughts will not be categorically readily indiscriminable to him (although they may be categorically *specially* indiscriminable to him). A colour-blind human will be categorically unable to readily discriminate between other states than humans with healthy vision. A colour-blind human does not have the general ability for colour-discrimination; for them more empirical information is categorically beyond what is readily available to them. One could set up a twin case for colour-blind humans where the twin case differs from the actual case only with respect to colour in a way that would be readily discriminable to humans

with healthy vision. For example, a twin case for people with red-green blindness could be such that in the twin case all the objects are green that are red in the actual case; such a case would not constitute a twin case to humans with healthy vision as the cases would be readily discriminable to them. What humans are categorically unable to do, other species may be generally unable to do and others able to do. Take the possible species with the general ability for molecular vision, for example. The classical Twin Earth scenario would have to take a different form for such a species since water and twin water are readily discriminable to them, if presented to them successively like in a switching case. If they are self-reflective, as we have been assuming, their water-thoughts and twin water-thoughts will be readily discriminable to them as well.

The categorical conception of an inability is therefore much weaker than the impersonal conception of an inability. The attribution of an impersonal inability to an individual is, as we know, not relative to any particular individual's set of specific or general abilities. It abstracts away from any feature of any of the individuals that have that ability, holding fixed nothing but the most general features of a case (e.g. natural laws) that would define some individual-independent, absolute limitation on an ability. Expectably, it is much harder to determine what those absolute limitations on an ability would be.

My suggestion is that individuals are generally unable to readily discriminate between their states throughout a sorites series of mental states but they are categorically unable to readily discriminate between their mental states in and out of a twin case. This is because twin cases all have the same following structure. The features that distinguish the twin case from the actual case are *categorically* beyond the sphere of readily discriminable facts of the individual cogniser in a context. The twin cases that are usually discussed are implicitly formulated against the backdrop of the set of general abilities of humans. To humans, a hidden difference in the molecular structure between two natural kinds, or in the way some expression is used in different linguistic communities, or between two numerically distinct but qualitatively identical objects is categorically beyond the sphere of readily discriminable facts in a

context in which we suppose that they are ignorant about these differences. In fact, quite a few things are categorically beyond the readily discriminable to humans in a context in which they are ignorant about them. Only new empirical evidence would make such differences discriminable to them. To other humans and other species in other contexts, other differences may be categorically beyond the readily discriminable, and twin cases for those individuals would have to take a different form. The central claim of state externalism is that, holding fixed the general abilities and the knowledge of a particular individual in a context, the individual's mental states may partly supervene on that which lies outside of the sphere of those facts that are readily discriminable to that individual in that context. This is the basis for the categorical epistemic definition of the debate I will propose.

4.3.1.2 Identifying, sorting, predicting

Externalism about mental states, according to the categorical epistemic definition that I propose, consists in claiming that even if some relevant mental states are categorically readily indiscriminable to an individual, they may be in distinct mental states. Here is the categorical epistemic definition of state externalism:

> *Categorical epistemic definition of state externalism*: a theory is state externalist if and only if, according to it, there is a case α that is such that S is *categorically* unable to readily discriminate the case α in which S is in M_α presented to S as "my current state of thinking that *p*" from an alternative case β in which S is in M_β presented to S either as "my earlier state of thinking that '*p*'" or as "my counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which S *categorically* does not readily know that it is not S's world), and yet $M_\alpha \neq M_\beta$.

State internalism is defined as the negation of state externalism:

> *Categorical epistemic definition of state internalism*: a theory is state internalist if and only if, according to it, if S is *categorically* unable to readily discriminate the case α in which S is in (the purely internal) M_α presented to S as "my current state of thinking that *p*" from an alternative case β in which S is in (the purely internal) M_β presented to S either as "my earlier state of thinking that '*p*'" or as "my counterpart's state of thinking that '*p*'" (where "counterpart" is defined as S's physical duplicate who lives in a world of which S *categorically* does not readily know that it is not S's world) at t, then $M_\alpha = M_\beta$ at t.

To get a clearer grip on how the categorical epistemic definition works, let us have a look at how it identifies, sorts, and predicts.

I will be able to state the substantial point of disagreement between state internalists and externalists more succinctly with the help of further terminology. Let's define "switching" as follows:

> *Switching*: A mental state switching occurred if and only if the following conditions are jointly satisfied: individual S is in M_α in case α and S is in M_β in case β, S is *categorically* unable to readily discriminate α from β (given the relevant modes of presentation), and $M_\alpha \neq M_\beta$.

The categorical epistemic definition of the state internalist/externalist debate identifies the possibility of mental state switching as the substantial point of disagreement between state internalists and externalists.

Importantly, I define internalism and externalism about mental *states*, i.e. the compound of an attitude and a corresponding content. Internalism about mental states rules out that the thought content or the attitude could switch, i.e. that the content or the attitude could be distinct between cases α and β, and yet the individual is categorically unable to discriminate between cases α and β. But the definition can be generalised to any mental condition, i.e. thought contents, attitudes, phenomenal states, specially accessible states

or any other condition, considered in isolation as well. A view is a version of state externalism about some mental condition if and only if switching with respect to that mental condition is possible. A view is state internalist about a mental condition if and only if it denies that switching with respect to that condition is possible.

The categorical epistemic definition defines the debate around the state externalist claim that mental state switching is possible. This can seem like a distortion of the dialectical situation between state internalists and externalists. Usually, the dialectic is considered to be as follows: state internalism is committed to a strong supervenience claim whereas state externalism merely denies state internalism. I believe that there has been a distortion of the debate insofar as the implications of the state externalist position have been underemphasised and those of the state internalist position made to look stronger than they are. The categorical epistemic definition nicely corrects that misbalance. Here is how.

If an individual's mental states partly supervene on external factors, their mental states will, in certain cases, vary in ways that the individual could only come to know about in the same way in which they come to know about anything that they are not already in a position to know: by more empirical evidence.[73] This is because the external, in some cases, extends beyond the sphere of the facts that are readily discriminable to an individual in a context. Fundamentally, state internalists and externalists disagree about how closely the mind is shaped by its surroundings. But an individual's surroundings extend beyond the readily discriminable. Switching is a direct consequence of that.

While switching is a direct consequence of state externalism, state internalism is not committed to implausibly strong views about the access individuals have to their mental states just in virtue of denying the possibility of switching. State internalists can accept many weaker forms of failures to readily discriminate between one's mental states, namely all those that result

[73] Note, importantly, that while mental state changes will be categorically readily indiscriminable in certain circumstances, the mental states may be *readily accessible* at any moment in virtue of being compatibilistically readily accessible. I will come back to this in a moment.

from an individual's specific or general inability to readily discriminate between their mental states in a context. The categorical epistemic definition of the state internalist/externalist debate thereby nicely highlights the difference between being a state externalist and accepting the non-luminosity of mental states.

A mental condition is luminous if and only if the individual is always in a position to know that they are in that mental condition if they are. Williamson has presented a well-known argument for the claim that no non-trivial mental condition is luminous (2000: Chap.4). The argument crucially relies on the possibility of devising a sorites series of any non-trivial mental condition of non-ideal cognisers. For any non-trivial mental condition of an individual with limited discriminative abilities, we may describe a series of incremental changes to the condition that add up to a big change. In between two consecutive cases, the mental condition must have changed, yet the change is too small for the individual to register. In those cases, the individual is in a position to know that they are in a mental condition only within a certain margin of error (see Williamson 2000: Chap.5). So, non-ideal cognisers are not always in a position to know that they are in a particular non-trivial mental condition if they are.[74] Williamson's non-luminosity argument crucially rests on the empirical fact that non-ideal cognitive systems have limited discriminative abilities, amongst others, with respect to their own mental states (see also Srinivasan 2015 for emphasis of this point).

Switching is different: switching is a mental state change that is readily indiscriminable to the individual but not explainable by any of the individual's shortcomings. So, it is *categorically* readily indiscriminable in my wording. Here is how Farkas puts the difference between state externalism and non-luminosity:

> "(…) phenomena like self-deception, difficulty of grasping complex ideas or the effects of strong emotional involvement suggest that such states [beliefs, desires or intentions] are often not known with first-person

[74] Note that the failure to readily know that one is in a mental condition that falls within the margins is discriminatory as well as compatibilist: the self-ascriptive belief is not safe if its content falls within the margins.

> authority. For reasons like this hardly anyone would want to maintain that we have unrestricted privileged access to all of our mental states. The striking feature of externalism is that it forces a limitation on privileged access which is fundamentally different in character: it arises with respect to the simplest current thoughts and experiences, and it is not explainable by these familiar facts of human psychology" (Farkas 2003: 203).

As was discussed in chapter 3, switching cases (or the risk of a switching case) sometimes occur according to state externalists. In switching cases, individuals lack discriminatory ready access to simple, occurrent mental states such as "I like prune tarts" or "this sushi roll looks tasty". Recall the circumstances in which you lack ready discriminatory access to the latter, for example, if state internalism is assumed. If a very similar looking sushi roll has taken the place of the one you were first looking at, your demonstrative thought will have changed its subject without your reckoning according to singular externalists. In such a case, you will not know what you are thinking a particular demonstrative thought instead of a relevant alternative demonstrative thought, assuming singular externalism.

That is what the categorical epistemic definition emphasises: state externalism is the view that one can think a determinately water-, or anxious depression-, or holy wisdom-, or prune-thought without being in a position to discriminate this thought from a relevant, alternative twin water-, or irritable depression-, or immaculate wisdom-, or plum-thought purely in virtue of the fact that one would need to find out more about one's environment in order to activate knowledge of distinctness. At the same time, that's the only case state internalists rule out *categorically*, if you will: it cannot be that an individual fails to readily know that they are a particular state instead of some relevant alternative state purely in virtue of the fact that they lack empirical evidence, absent any shortcoming on their part. If an individual hasn't missed something that they wouldn't have missed if they were better cognisers, there was nothing to miss; so if they are categorically unable to readily discriminate between two states, they must be in the same state according to state internalists.

This fault line is exactly what one would expect from a debate on the extent to which an individual's mental states are shaped by the individual's

surroundings. The substantial disagreement that the categorical epistemic definition identifies—*is mental state switching possible?*—is a worthy point of debate.

Next, let's see how the categorical epistemic definition sorts different views. The important taxonomical criterion is that any view according to which switching is possible is state externalist. From the last chapter, we know that switching occurs for any sort of externalism about thought contents as well as externalism about attitudes. So, natural kind externalism, social externalism, singular externalism and attitude externalism are correctly classified as state externalist theories. The categorical definition also correctly classifies phenomenal externalism and externalism about specially accessible states. If, for example, in the switching case to the classical Twin Earth scenario, Oscar is categorically unable to readily discriminate between his water-thoughts and his twin water-thoughts, he will be categorically unable to readily discriminate between water-seemings and twin water-seemings.

Note that it does not matter whether the natural kind is located only within the boundaries of the individual's body, or dependent on material or immaterial substances located outside of the individual or not located at all. A view according to which Oscar and Twin Oscar are in distinct, yet readily indiscriminable states in the anxious depression/irritable depression case, independently of the location of the natural kind, is correctly classified as a form of state externalism. Whether located or not, the Holy Spirit and the Virgin Mary will sometimes change in ways that an individual is categorically unable to readily discriminate, i.e. is unable to notice without further *spiritual* evidence about the states of the Virgin Mary and the Holy Spirit, not owing to a limitation in their abilities.

Any view according to which switching is impossible will be classified as state internalist, irrespective of the specific state internalist's views on the mind-body or the intentional-phenomenal relationship. Those dualist views that reject the view that a mental state change that is categorically readily discriminable could take place will be classified as state internalist. Take Cartesian dualism. Whatever Descartes' view on the metaphysics of mental

states, his views on the epistemology of mental states are such as to exclude the possibility of switching.

The categorical epistemic definition states a clear recipe for formulating an externalist view on any mental condition: show that switching is possible with respect to that condition. This is the recipe that any of the views that extend the scope of externalism about the mental follow: phenomenal externalists show that switching is possible not only with respect to an individual's intentional states but also with respect to their phenomenal states; special access externalists show that switching is possible not only with respect to an individual's intentional and phenomenal states but also with respect to their specially accessible states. The categorical epistemic definition correctly classifies all views and cases encountered in this chapter.

Let's come to the predictions the categorical epistemic definition makes. According to the categorical epistemic definition, it is an analytic consequence of state externalism that individuals lack a certain kind of access to their externally-individuated mental states in certain circumstances. Brie Gertler considers this to be a major problem of the epistemic definition. She writes:

> "The Epistemic Criterion has some problematic consequences. First, it ensures that externalism is incompatible with privileged first-person access, as a definitional matter. The Epistemic Criterion glosses externalism as the claim that content properties do not supervene on (and hence, aren't identical to) properties to which the thinker enjoys privileged access. Farkas embraces this consequence, saying that "one way to sum up my proposal is to say that externalism is a thesis about the nature of our access to our thoughts" (Farkas 2003: 204). While most externalists concede that their view initially appears incompatible with privileged access, most also maintain that these are ultimately compatible. Regardless of whether compatibilism is true, the controversy surrounding this issue casts doubt on the idea that incompatibility with privileged access is a simple analytic consequence of externalism" (Gertler 2012: 61).

I agree with Farkas as quoted here: the state internalist/externalist debate is fundamentally about the limits of self-knowledge. However, it is not fundamentally about the limits of self-knowledge *tout court* because, I agree

with Gertler, the question of whether compatibilism is true or not is not an analytic one. Recall, compatibilists believe that (self-reflective) individuals always (even in switching cases) have ready access to certain kinds of external mental states in virtue of the fact that they can self-ascribe them in a particular way (for detailed discussion see §3.3.2). Instead, the state internalist/ externalist debate is fundamentally about the limits of *discriminatory* self-knowledge. This is because, if state externalism about some mental condition is assumed, in switching circumstances, the individual will be categorically unable to readily discriminate their occurrent external state from some relevant alternative state. That this would be an analytic consequence of state externalism seems to me unsurprising. Again, state internalists and externalists disagree about the extent to which an individual's environment, whether natural, linguistic, objectual or phenomenal, shapes the individual's mind. If the nature of an individual's mind is closely tied to the nature of their environment, it is to be expected that the epistemology of an individual's mind will as well.

4.3.2 Substantial metaphysical accounts

We should read the physical and phenomenal definitions not as definitions but as substantial state internalist accounts of how particular individual's mental states are individuated (**§ 4.3.2.1**). Read in that way, they are explanatorily powerful theories that help to prevent the explanatory gaps that would be caused by a misunderstood epistemic definition of state internalism (**§ 4.3.2.2).**

4.3.2.1 Substantial state internalist accounts

In contrast to the definition of a debate, the subject matter of substantial accounts is a specific phenomenon, such as the nature of mental states, of which a complete account is given. Substantial accounts offer explanatorily powerful theories partly in virtue of committing to negotiable assumptions that will not be shared by everyone like-minded.

A materialistically-minded state internalist may, for example, hold that an individual's mental states supervene on the individual's physical states. According to such an account, if two individuals are in the same physical states, they are in the same mental state. We can define:

> *Materialist state internalism*: an individual's purely internal mental states at t supervene on the individual's physical states at t.

A materialist state internalist account commits to a certain view on the mind-body relationship. It will assume that the right psycho-physical laws exist to ground the metaphysics of an individual's mental states in the individual's physical states. According to materialist state internalists, if an individual is in the same physical states in and out of a twin case, they are in the same mental state in and out of a twin case; if an individual is in distinct mental states throughout a sorites series of mental states, they are in distinct physical states throughout the sorites series. Naturally, a materialist state internalist is inconsistent with dualist versions of state internalism.

A dualist state internalist, in contrast, may hold that an individual's mental states supervene on the individual's phenomenal states:[75]

> *Dualist-phenomenalist state internalism*: an individual's purely internal mental states at t supervene on the individual's phenomenal states at t.

A dualist-phenomenalist commits to a certain view on the intentional-phenomenal relationship, namely that intentional content supervenes on the phenomenal. Depending on the scope of the view, it will commit to a phenomenalist account of other mental phenomena as well. It will assume a specific internalist view on phenomenal states that is, of course, inconsistent with other views on the nature of phenomenal states such as phenomenal

[75] Not every phenomenalist account of mental phenomena is dualist and not every dualist account is phenomenalist. Here it is assumed that a dualist account of phenomenal states is given, i.e. an account according to which phenomenal states belong to an essentially different kind than other natural mental kinds.

externalism (see Farkas 2008: 124f.). The metaphysics of an individual's mental states, according to such a view, will be grounded in the individual's phenomenal states. According to dualist-phenomenalist state internalism, if an individual is in the same phenomenal states in and out of a twin case, they are in the same mental state in and out of a twin case; if an individual is in distinct mental states throughout a sorites series, they are in distinct phenomenal states throughout the series.

The materialist state internalist and the dualist-phenomenalist state internalist accounts are accurately sorted as state internalist according to the categorical epistemic definition. This is because there are no cases where an individual is categorically unable to readily discriminate between the relevant mental states and they are not in the same mental state, on either of the materialist or the dualist state internalist account. No doubt, a non-ideal individual may mistake distinct mental states to be same (they are actually in distinct physical/phenomenal states in the relevant cases) because of tiredness, distraction, forgetfulness, or inbuilt limitations. There may also be cases where a non-ideal individual may mistake the same mental state to be distinct (they are actually in the same physical/phenomenal states in the relevant cases), for example, in virtue of some serious confusion. But once we abstract away from such temporary or permanent obstructions to an individual's inability to readily discriminate between their states, there does not seem to be space for mistakes on a state internalist account. This is because, if a mental state change occurs, necessarily involving a change in the individual's physical or phenomenal states according to these accounts, and the change remains unnoticed by the individual, this must be in virtue of limitations regarding the individual's discriminative abilities. Put differently, the mental state changes that supervene on changes to an individual's physical or phenomenal states are never categorically beyond the individual's sphere of readily discriminable facts. It follows that if an individual is categorically unable to readily discriminate between their states, this must be because they are in the same state according to these state internalist views. The materialist and dualist state internalist views both come out as state internalist on the categorical epistemic definition of state internalism.

4.3.2.2 Closing the gaps

If we read the categorical epistemic definition of state internalism as stating state internalist conditions for mental identity in certain circumstances, amongst others in twin cases, it would result in the same explanatory gaps as the impersonal epistemic definition of state internalism (if *it* is read as stating state internalist conditions for mental identity).

Take the classic Twin Earth scenario. Oscar's water-thoughts are categorically readily indiscriminable from Twin Oscar's twin water-thoughts. State internalists hold that Oscar and Twin Oscar are necessarily in the same mental state. But it seems clear that, if that is correct, it is not *in virtue* of the fact that Oscar's ideal counterpart is unable to readily discriminate between their (own) states in the Twin Earth scenario. This is the resurgence of the first explanatory gap for the categorical epistemic definition.

Again, like the impersonal epistemic definition, the categorical definition of state internalism would also seem to lack the resources to provide an explanation for why the relevant mental states are categorically readily indiscriminable. Once one abstracts away from the individual's temporary and permanent limitations on their discriminative abilities, what, if not some robust properties of the mental states, could explain why they are categorically readily indiscriminable? But the categorical epistemic definition does not mention any such robust properties. That is the return of the second explanatory gap.

The gaps are prevented from opening once we take substantial state internalist accounts, and those only, to present state internalist conditions for mental identity and distinctness.

How is the fact that the ideal counterpart of some individual is unable to readily discriminate between their (own) relevant mental states going to explain that the non-ideal individual is in the same mental state in certain circumstances? The short answer is: it is not going to. It would be hopeless to derive sufficiency conditions for state identity for non-ideal individuals from the categorical epistemic definition of state internalism: how could the metaphysics of particular individual's mental states be explained by the

epistemology of their ideal counterpart's mental states? Trying to appeal to the absolute limitations on an ability, as Parrot and Gomes suggest we should do makes matters worse, if anything: it is even less clear what the metaphysics of particular individual's mental states, being the minds of individuals with particular permanent and temporary features, have got to do with the limitations of a possible epistemic faculty that we identify by abstracting away from any particular feature of the individuals who have that faculty. If the categorical epistemic definition was the theoretical repertoire for building an account of the metaphysics of mental states of particular individuals that state internalists have at their disposal, of course they would be in trouble.

The appeal to the ideal counterpart of an individual in a context merely fulfils the function of identifying the sort of mistake that, according to state internalists, individuals could not make about their mental states: a mistake that is due entirely to empirical ignorance and to none of the individual's shortcomings. Unlike Parrott and Gomes' impersonal definition, the categorical definition identifies a *good* epistemic criterion for the identity of a particular individual's purely internal mental states in a context accoridng to state internalism because it is implicitly defined against the backdrop of the sphere of readily discriminable facts to that individual in that context. It says that any variation of the individual's mental states in that context that would be caused by factors outside of the sphere of facts readily discriminable by that individual in that context are ruled out from happening by state internalism. It is the fact that the ideal counterpart of that individual would have detected any variation within that sphere that sees to it.

Instead it is substantial state internalist accounts that give accounts of the metaphysics of particular, non-ideal individuals' mental states. According to substantial state internalist accounts, particular individuals' mental states will, as one would expect, vary with not only the permanent but also temporary features of the particular individual in a context: a colour-blind human will be in different mental states from a human with healthy vision in a context because they will be in different physical and/or phenomenal states; the mental states of a sommelier at a wine tasting will differ from those of a layman because the sommelier will be in different physical and/or phenomenal

states at the tasting; and the mental states of a sober sommelier will differ from their mental states when drunk because they will be in different physical and/or phenomenal states in that context.[76] And, importantly, in a twin case, the particular individual is in the same mental state in virtue of being in the same physical and/or phenomenal states in and out of the twin case according to such state internalist accounts.

How are state internalist accounts going to explain that the relevant mental states are readily indiscriminable to the individual's counterpart if not in virtue of some robust properties that those states share? The short answer is: precisely like that. Substantial state internalist accounts will explain the epistemic features of the mental states of any individual in terms of the individual's physical or phenomenal states. The epistemic features of the ideal counterpart's mental states are no exception to this. Different state internalist accounts will ground the ready indiscriminability of the ideal counterpart's mental states in different metaphysical facts about their mental states: materialist state internalists may explain the epistemic fact by reference to the sameness of physical states or of the specially accessible properties; phenomenalist-dualist by reference to the sameness of phenomenal properties; Cartesian dualists by reference to the sameness of immaterial substances. State internalists do not need to accept mysterious, ungrounded epistemic facts because the ready indiscriminability of the ideal counterpart's mental states will be grounded in some underlying metaphysical identity.

Note as well that, just like state internalists, many state externalists will prefer metaphysically robust, in contrast to purely epistemic, explanations of the identity and distinctness of an (ideal or non-ideal) individual's mental states. State externalists may refer to contextually salient, metaphysical features that explain why two utterly distinct mental states are categorically readily indiscriminable. Martin suggests, for example, that the same underlying functional structures may suffice to explain why states of a different kind are readily indiscriminable to an individual (2006: ftn. 12). State

[76] For the same reason, it will not be an issue for substantial state internalists to provide an account of the identity and distinctness of non-reflective animals' mental states. Such accounts will not have anything to do with a capacity for ready discrimination between mental states, but with physical and phenomenal states that non-reflective animals are of course just as much in as we are, according to state internalists.

externalists will just deny that there is an explanation in terms of a shared property available across all contexts.

4.3.3 Diagnosis and treatment: the overall picture

The distinction between the definition of the state internalist/externalist debate and substantial accounts therein allows one to define the debate, and specifically the state internalist position in a certain way, while avoiding the explanatory gaps that other epistemic definitions would face. I want to close by giving a diagnosis of why the distinction has been missed in the debate on the definition of the state internalist/externalist debate specifically (**§4.3.3.1**). We will see that once the combination of the categorical epistemic definition and substantial accounts is applied, a number of things fall neatly into place (**§4.3.3.2**).

4.3.3.1 Diagnosis

The reason why the distinction between the definition of the state internalist/ externalist debate and substantial accounts, and specifically substantial state internalist accounts, within the debate has been missed, stems from a mistreatment of twin cases.

As we have seen throughout the chapters, twin cases play a crucial role in setting up the debate between state internalists and state externalists: state externalist positions were developed with the help of twin cases, any state externalist view can be formulated with the help of a twin case, and, crucially, they divide state internalists and externalists. Twin cases are set up in a certain way, namely by holding certain facts about an individual fixed across the actual and the counterfactual environment and by varying certain environmental factors. It is therefore a short step to the following inference: whatever the facts are that one needs to hold fixed across environments, these facts are sufficient for mental state identity according to state internalists and are not sufficient according to state externalists.

In the literature, one can therefore see two questions being commonly run together: "How is the state internalist/externalist debate defined?" and "What is the correct account of 'internal sameness' or of the 'twin relation' between two individuals?". In her paper "What is Externalism?", Farkas writes "What we need then is a characterization of a relation between the Twins which (…) establishes an identity or equivalence between the Twins in some respect (2003: 184). In her book, she adds

> "The question is: how should we understand the relation constitutive of the Twin situations? This is important, first, because this relation is the basis of the internalism/externalism controversy. Internalists say that subjects in the Twin situations have the same mental features; externalists deny this" (Farkas 2008: 86).

Parrott and Gomes agree with Farkas' on the methodology for finding the correct definition of the state internalist/externalist debate. They write:

> "Using this framework we can distinguish different accounts of the internalism/externalism distinction in terms of the requirements which have to hold in order for two agents to be internally the same (…) differing accounts of internal sameness (…) will present us with different ways of understanding the distinction between internalism and externalism" (2021: 316).

The crucial background assumption is that state internalists and externalists agree on an account of internal sameness but disagree on the implications of internal sameness for the metaphysics of mental states of the relevant individuals: if two individuals are internally the same, they must be in the same mental state according to state internalists; and may be in distinct mental states according to state externalists. The crucial result of this methodology is that the candidate definitions would seem to kill two birds with one stone: they define the state internalist/externalist debate as well as, or rather, by means of, stating sufficiency conditions for mental identity according to state internalism.

But at least some state internalists and externalists do not agree on any account of internal sameness. The dynamic that one can observe unfolding with Farkas' and Parrott and Gomes' methodology is the following. Whatever

property state internalists would propose as an account of internal sameness is rejected by some state externalists: externalists about thought content reject the view that individuals in and out of twin cases are internally the same in virtue of thinking thoughts with the same contents; phenomenal externalists reject the view that individuals in and out of twin cases are internally the same in virtue of being in the same phenomenal states; externalists about self-knowledge reject that individuals in and out of twin cases are internally the same in virtue of being in the same specially accessible states. What is left is the epistemic fact that an individual's mental states are categorically readily indiscriminable to the individual in and out of twin cases.

State internalism is left in the desolate place of trying to extract plausible sufficiency conditions for mental identity out of the fact that the relevant mental states are categorically readily indiscriminable to that individual in that context. We know what follows: multiple explanatory gaps.

But twin cases have first and foremost been designed by state externalists to introduce different versions of state externalism as a challenge to state internalism. So while we can expect twin cases to provide essential insights into the fundamental point of disagreement between state internalists and externalists, it would be surprising if they also provided substantial state internalist accounts of mental identity. If it is surprising, it is likely that they don't.

4.3.3.2 Treatment

The mental states of an individual in and out of a twin case are categorically readily indiscriminable to that individual. The distinction between the definition of the state internalist/externalist debate and substantial accounts, and importantly substantial state internalist accounts within the debate, clarifies what the appeal to an individual's categorical ready indiscriminability does and what it does not do. Let us start with the latter.

The fact that some relevant mental states of an individual are categorically readily indiscriminable to that individual in and out of a twin case is not the last bit of shared mental space into which increasingly

extravagant forms of state externalism force state internalism, and which would then have to suffice, however narrow it is, to provide a substantial state internalist account of mental sameness. Why would a phenomenalist state internalist care about the fact that some (crazy) phenomenal externalists believe that phenomenal states are also externally-individuated? The right answer is, they wouldn't. Phenomenalist state internalists may of course appeal to phenomenal states in an account of mental identity, relying on a decidedly internalist conception of phenomenal states. It is not twin cases but substantial state internalist accounts that give accounts of internal sameness: they specify the conditions under which two individuals are internally the same according to state internalists, such that it is entailed that they are in the same purely internal mental states.

What the fact that an individual's mental states are categorically readily indiscriminable in and out of a twin case does tell us is what is common to any state internalist account as opposed to any state externalist account of mental states (as well as any other mental condition). By virtue of their view, state internalists rule out that individuals may be mistaken about the fact that they are in a particular occurrent mental state rather than in some particular alternative mental state purely out of empirical ignorance, and not owing to shortcomings in any of their discriminative abilities. But state internalists may accept that individuals lack discriminatory ready access to their mental states in virtue of the fact that they are limited cognisers. By virtue of their view, in turn, state externalists allow that the mental states of a particular individual in a particular context may vary with features that are categorically beyond the sphere of readily discriminable facts to this individual in that context. Pinpointing the essential commitment common to any state internalist view as opposed to any state externalist view is the only task that the categorical epistemic definition fulfils.

What emerges is the following, slightly paradoxical relation of substantial state internalist and substantial state externalist accounts to the categorical epistemic definition of the debate between them. The definition groups together theories according to which the fact that some states are categorically readily indiscriminable is irrelevant with regard to the

metaphysics of states, although irrelevant in different ways. To state internalists, categorical ready indiscriminability is the trivial effect of state identity; to state externalists, categorical ready indiscriminability is a collateral effect unrelated to the metaphysics of external mental states.

4.4 Conclusion

The debate on the consistency of hybrid access internalism tends to revolve around the question of whether we have special access to our external mental states. This is because the special, introspective, non-empirical access we have to our own minds has been put at the centre of the debate on hybrid access internalism from two independent directions.

In epistemology, access internalist theories have emphasised the importance of special access because special access would seem to provide the crucial link between the two central intuitions that access internalists like to emphasise about justification: the access motivation and the Equal Justification Thesis. If the access to one's evidence and to facts about one's evidence on which the justification of one's beliefs supervenes is special access, then individuals in really bad cases may have justification to believe the same propositions as individuals in the good case to the same extent. At the same time, in the philosophy of mind, theorists have been worried that, if state externalism is true, individuals would lack special access to their external mental states. So if individuals lack special access to their external mental states, this would not only seem upsetting in itself but it also undermines both motivations behind access internalism in one blow.

One of the overarching themes of this thesis has been to remove special access from the centre of the debate on hybrid access internalism, and to put ready access at its place. If it is ready access that access internalists want and that state externalists threaten to take away, this has at least three non-radical yet remarkable effects on the debate on the consistency of hybrid access internalism.

First, the internalist positions discussed are more plausible. On the back of very substantial assumptions about the kind of properties that individuals have special access to, epistemic internalists and state internalists alike tend to emphasise phenomenal properties, appearances and seemings to be the fundamental stuff that purely internal mental states are made off and on which justification supervenes. But these properties are hard to grasp and do not seem to play a role in our daily routines of thinking and justifying ourselves to others (and ourselves). At the same time, things that are easy to grasp (sometimes literally) and important to our daily practices because they are readily available are excluded from being relevant to internalist philosophies of mind or justification theories because we cannot know them in a special or reflective but merely in an empirical way. This leaves internalist theories, I think, in an extremely implausible shape.

Epistemic internalists and state internalists are often motivated by the aim of construing a mental reality that is independent of the individual's environment and protected from the mistakes they make about it. But, and this is the second effect that focussing on ready access has, by pushing the boundary of the internal/external out of the mind and to the limits of the readily discriminable, internalists do not have to give up much of their internalism except for being prisoners of their own minds. Of course, if the boundary of the internal/external runs around the readily discriminable/non-discriminable, what mental states an individual is in and what they have justification to believe is no longer independent of the individual's environment. But internalists do not have to accept that individuals could make just the same mistakes about their mental states or their evidence that they make about any empirical fact. Only state externalists have to accept that individuals may make mistakes about their minds purely out of empirical ignorance and not owing to a shortcoming in themselves.

In virtue of the fact that very similar conflicts arise when state internalist positions are formulated with the more intuitive and more permissive notion of ready access shows that the special way in which we know our minds (if it is so special) is dispensable in understanding the tension between access internalism and state externalism. Externalism about any

mental condition is, by definition, a view that limits the kind of ready access that we have to that mental condition in certain circumstances just by virtue of how externalism considers our surroundings to shape our minds. So the third effect is that, not only are the views discussed more plausible, we also get a better understanding of what they are and how they conflict.

I have defended the view that state internalism/externalism is defined by reference to the possibility of mental state switching. I want to suggest that epistemic internalists and externalists disagree about the possibility of "justification switching", i.e. whether the justification of an individual's beliefs could change in ways that are categorically indiscriminable to the individual. Such a discrimination definition of access internalism would look as follows:

> *Categorical epistemic definition of access internalism*: a theory is access internalist if and only if, according to it, if S is *categorically* unable to readily discriminate the case α in which S has justification to believe B_α to extent α, from an alternative case β in which S has justification to believe B_β to extent β, then the extent to which S has justification to believe B_α in α = the extent to which S has justification to believe B_β in β.

Epistemic externalism would be the view that even if S is categorically unable to readily discriminate between cases α and β, S may have justification to believe B_α and B_β to different extents. Note that if $B_\alpha \neq B_\beta$, then the categorical epistemic definition of access internalism is equivalent to the Indiscriminable Justification Thesis suggested in chapter 2. So, the categorical inability to readily discriminate between cases may provide the needed qualification of the Indiscriminability Justification Thesis as well.

I do not have the space to defend the application of the framework developed in this chapter to the internalism/externalism debate in epistemology. But I want to highlight where we would gain new clarity on the challenges to the consistency of state externalism and epistemic internalism if the framework did apply in epistemology.

The central question of the consistency of hybrid epistemic internalism would then be: *Does the possibility of mental state switching entail the possibility of justification switching?*

A cursory look says that it does for hybrid mentalism: if your mental states have been switched, and justification supervenes on your mental states, then whether you have justification to believe some proposition or not is likely to have switched as well. If justification switching is furthermore definitory of epistemic externalist views, hybrid mentalism comes out as an epistemic externalist view.

For hybrid access internalism, the situation is more complex: if your mental states have been switched, and justification supervenes on those facts about your evidence to which you have ready access, then the question is whether you have ready access to the relevant facts about your evidence. Given the findings in chapter 3, you are unlikely to have discriminatory ready access to the facts about your evidence in switching cases, but there is still compatibilist ready access. It will be interesting to explore whether mental state switching entails justification switching, if we assume compatibilist access to our external states in a hybrid access internalist theory.

The picture that emerges is that of a unified way of understanding the internalism/externalism debates in philosophy of mind and epistemology. Both debates are concerned with the relationship between certain properties of philosophical relevance—the nature of mental states, on the one hand, and the nature of justification, on the other hand—and those parts of an individual's environment that are beyond the individual's readily discriminable ken.

Bibliography

Adams, F.; Drebushenko, D.; Fuller, G. & Stecker, R. (1990). Narrow content: Fodor's folly. Mind and Language 5 (3):213-29.

Alston, W. (1971). Varieties of privileged access. American Philosophical Quarterly 8: 223–41.

Alston, W. (1986). Internalism and externalism in epistemology. Philosophical Topics, 14, 179–221.

Alston, William P. (1988). An internalist externalism. Synthese 74 (3):265 - 283.

Antony, L. (2007). Everybody has got it: A defense of non-reductive materialism. In Brian P. McLaughlin & Jonathan D. Cohen (eds.), Contemporary Debates in Philosophy of Mind. Blackwell.

Armstrong, D. (1968). A Materialist Theory of the Mind. London: Routledge.

Armstrong, D.M. (1973). Belief, Truth, and Knowledge. Cambridge University Press: London.

Armstrong, D. M. (1981). The Nature of Mind and Other Essays. Ithaca, NY: Cornell University Press.

Audi, R. (2001). An internalist theory of normative grounds. Philosophical Topics, 29, 19–46.

Audi, Robert (2003). Contemporary Modest Foundationalism in Louis J. Pojman (ed.) The Theory of Knowledge: Classical and Contemporary Readings. (Belmont, CA: Wadsworth).

Bear, A., & Knobe, J. (2017). Normality: Part Descriptive, Part Prescriptive. Cognition, 167, 25–37.

Bergmann, M. (2006). Justification without awareness. Oxford: Clarendon Press.

Bernecker, S. & Dretske, F. (eds.) (2000). Knowledge: Readings in Contemporary Epistemology. Oxford University Press

Berker, S. (2008). Luminosity Regained. Philosophers' Imprint, 8(2): 1–22.

Blackburn, S. (1984). The individual strikes back. Synthese 58 (March):281-302.

Block, N., (1986). Advertisement for a Semantics for Psychology. Midwest Studies in Philosophy, 10: 615–678. Reprinted in Stephen P. Stich and Ted A. Warfield, eds., Mental Representation: A Reader, Oxford: Blackwell, 1994.

Block, Ned, 1990, Inverted Earth, in Philosophical Perspectives 4, James Tomberlin (ed.), Atascadero, CA: Ridgeview Press, 52–79.

Block, N.(1995). On a Confusion about a Function of Consciousness. Behavioral and Brain Sciences,18 (2): 227-287.

Block, N. (1996). Mental paint and mental latex. In E. Villanueva (Ed.), Philosophical issues (Vol. 7). Atascadero, CA: Ridgeview Publishing.

Block, N. (1997). Anti-Reductionism Slaps Back. Noûs 31 (s11):107-132.

Block, N. (2011). Perceptual consciousness overflows cognitive access. Trends in Cognitive Sciences 15 (12):567-575.

Block, N. & Stalnaker, R. (1999). Conceptual analysis, dualism, and the explanatory gap. Philosophical Review 108 (1):1-46.

Boghossian, P. (1989). Content and self-knowledge. Philosophical Topics, 17, 5–26.

Boghossian, P. (1992). Externalism and inference. Philosophical Issues, 2, 11–28.

Boghossian, P. (1994). The transparency of mental content. Philosophical Perspectives, 8, 33–50.

Boghossian, P. (1997). What the externalist can know a priori. Proceedings of the Aristotelian Society, 97, 161–175.

BonJour, L. (1985). The structure of empirical knowledge. Cambridge, MA: Harvard University Press.

BonJour, L. (1992/2010). Recent Work on the Internalism–Externalism Controversy. In J. Dancy & E. Sosa (Eds.),. A companion to epistemology (pp. 132–136). Oxford: Blackwell.

BonJour, L. (2002). Internalism and externalism. In Paul K. Moser (ed.), The Oxford Handbook of Epistemology. Oxford University Press. pp. 234–264.

Brown, J. (1995). The incompatibility of anti-individualism and privileged access. Analysis, 55(3),149–156.

Brown, J.(1998). Natural Kind Terms and Recognitional Capacities, Mind 107, pp. 275–303.

Brown, J. (2004). Anti-individualism and knowledge. Cambridge, MA: MIT Press.

Brown, J. (2007). Externalism in mind and epistemology. In S. Goldberg (Ed.), Internalism and externalism in semantics and epistemology (pp. 13–34). Oxford: Oxford University Press.

Brueckner, A. (1990). Scepticism about knowledge of content. Mind 99 (395):447-51.

Brueckner, A. (2002). The consistency of content-externalism and justification-internalism. Australasian Journal of Philosophy, 80(4), 512–515.

Burge, T. (1979). Individualism and the Mental. Midwest Studies in Philosophy, 4, 73–121.

Burge, T. (1982a). Two Thought Experiments Reviewed. Notre Dame Journal of Formal Logic 23.3, 284-293.

Burge, T. (1982b). Other Bodies. In Woodfield, Andrew, ed., Thought and Object: Essays on Intentionality. Oxford: Oxford University Press, 97-121.

Burge, T. (1986). Individualism and psychology. Philosophical Review, 95, 3–45.

Burge, T. (1988). Individualism and self-knowledge. Journal of Philosophy, 85, 649–663.

Burge, T. (1989). Individuation and causation in psychology. Pacific Philosophical Quarterly 707 (4):303-22.

Burge, T. (1998). Memory and self-knowledge. In P. Ludlow & N. Martin (Eds.), Externalism and self-knowledge (pp. 351–370). Stanford: CSLI Publications.

Byrne, A. (2012): Knowing what I see. Introspection and Consciousness, eds. D. Smithies and D. Stoljar, OUP 2012.

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

Byrne, A. & Tye, M. (2006). Qualia ain't in the head. Noûs 40 (2):241-255.

Callahan, L. (2020). Perception, discrimination, and knowledge. Philosophical Issues 30 (1):39-53.

Campbell, J. (1987). Is Sense Transparent? Proceedings of the Aristotelian Society, 61, 273–292.

Campbell, J. (2006). Reference and Consciousness. Philosophy and Phenomenological Research 72 (2):490-494.

Carruthers, P. (2011). The Opacity of Mind: An Integrative Theory of Self-Knowledge. Oxford: Oxford University Press.

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Chalmers, D. (2001). The nature of epistemic space. Reprinted in A. Egan and B. Weatherson, (eds.) Epistemic Modality. Oxford University Press, 2010.

Chalmers, D. (2002). The components of content. In David J. Chalmers (ed.), Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press.

Chalmers, D. (2006), Two-Dimensional Semantics, in Oxford Handbook of Philosophy of Language, E. Lepore and B. Smith (eds.), Oxford: Oxford University Press, pp. 575–606.

Chalmers, D. (2012). Constructing the World. Oxford University Press.

Chalmers, D. (2019) Review of Juhani Yli-Vakkuri and John Hawthorne, Narrow Content, Oxford University Press, 2018. https://ndpr.nd.edu/news/narrow-content/.

Chase, J. (2001). Is externalism about content consistent with internalism about justification? Australian Journal of Philosophy, 79(2), 227–246.

Chisholm, R. (1977). The theory of knowledge. Eaglewood Cliffs, NJ: Prentice Hall.

Chuard, P. & Southwood, N. (2009): Epistemic Norms without Voluntary Control in Noûs 43 (4): 599-632.

Churchland, P. (1970). The Logical Character of Action Explanations. Philosophical Review 79: 214–236.

Cohen, S. (1984). Justification and Truth. Philosophical Studies 46 (3):279--95.

Conee, E. (2007). Externally enhanced internalism. In S. Goldberg (Ed.), Internalism and externalism in semantics and epistemology (pp. 51–67). Oxford: Oxford University Press.

Conee, E. & Feldman, R. (2004). Internalism defended. Reprinted in Evidentialism: Essays in Epistemology. Oxford, England: Oxford University Press.

Crane, T. (1991). All the Difference in the World. Philosophical Quarterly 41 (162):1-25.

Dancy, J. (1985) An Introduction to Contemporary Epistemology, London: Oxford University Press.

Davidson, D.(1970). Mental Events. In L. Foster & J. W. Swanson (eds.), Essays on Actions and Events. Clarendon Press. pp. 207-224.

Davidson, D. (1984). First person authority. Dialectica 38: 101–11. Page reference to the reprint in Davidson 2001

Davidson, D. (1987), Knowing One's Own Mind, in Proceedings and Addresses of the American Philosophical Association, 61: 441–58.

Davies, M. (1998). Externalism, Architecturalism and Epistemic Warrant in Wright-Smith-MacDonald 1998, 321–361.

Davies, M. (1993). Aims and claims of externalist arguments. Philosophical Issues 4:227-249.

DeRose K. (1995). Solving the Skeptical Problem, The Philosophical Review, 104(1), 1–52.

Descartes, R. ([1641]/1998). Meditationes de prima philosophia/ Meditations and Other Metaphysical Writings, trans. Desmond M. Clarke. London: Penguin.

Dogramaci, S. (2012). Reverse Engineering Epistemic Evaluations. Philosophy and Phenomenological Research 84 (3):513-530.

Dokic, J. and Égré, P. (2009). Margin for Error and the Transparency of Knowledge. Synthese, 166(1), pp. 1–20.

Dretske, F. (1981). Scepticism: A Critical Appraisal. Philosophical Topics 12 (2):299-303.

Dretske, F. (1995). Naturalizing the Mind. MIT Press.

Dummett, M. (1970): 'Wang's Paradox', reprinted in his Truth and Other Enigmas, London: Duckworth, 1978, pp 248-68.

Ellis, J. (2010). Phenomenal character, phenomenal concepts, and externalism. Philosophical Studies 147 (2):273 - 299.

Evans, G. (1982). The varieties of reference. Oxford: Oxford University Press.

Falvey, K., & Owens, J. (1994). Externalism, self-knowledge, and scepticism. Philosophical Review, 103, 107–137.

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

Farkas, K. (2006).Indiscriminability and the sameness of appearance. Proceedings of the Aristotelian Society 106 (2):39-59.

Farkas, K. (2008a), “Phenomenal intentionality without compromise”, The Monist, 91(2): 273–93.

Farkas, K. (2008b). The Subject’s Point of View. Oxford University Press, Oxford

Feldman, R. (1985). Reliability and Justification. The Monist 68 (2):159-174.

Feldman, R. & Conee, E. (2001). Internalism Defended. American Philosophical Quarterly 38 (1):1 - 18.

Fernández, J. (2013). Transparent Minds: A Study of Self-Knowledge. Oxford University Press.

Fiddick, L., Cosmides, L. & Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. Cognition 77 (1):1-79.

Fodor, J. (1979). Methodological solipsism considered as a research strategy in cognitive psychology. Behavioral and Brain Sciences 3 (1):63-73.

Fodor, J. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press, Cambridge, Mass.

Fodor, J. (1997). Special sciences: Still autonomous after all these years. Philosophical Perspectives 11:149-63.

Fodor, J. & Davies, M. (1986). Individualism and Supervenience. Proceedings of the Aristotelian Society, Supplementary Volumes 60:235-283.

Fratantonio, G. (forthcoming). Evidential Internalism and Evidential Externalism. In Maria Lasonen-Aarnio & Clayton M. Littlejohn (eds.), The Routledge Handbook for The Philosophy of Evidence.

Frege, Gottlob ([1892] 1948). "Sense and Reference." *The Philosophical Review* 57.3: 209-30.

Fricker, E. (2009). Is knowing a state of mind? The case against. In Duncan Pritchard & Patrick Greenough (eds.), Williamson on Knowledge. Oxford: Oxford University Press.

Fumerton, R. (1995). Metaepistemology and Skepticism, Lanham, MA: Rowman and Littlefield.

Gaukroger, C. (2017). Why broad content can’t influence behaviour. Synthese 194 (8):3005–3020.

Gerken, M. (2008). Is internalism about knowledge consistent with content externalism? Philosophia, 36, 87–96.

Gertler, B. (2012). Understanding the internalism-externalism debate: What is the boundary of the thinker? Philosophical Perspectives 26 (1): 51-75.

Gibbons, J. (1996). Externalism and knowledge of content. Philosophical Review, 105, 287–310.

Gibbons, John (2006). Access externalism. Mind 115 (457):19-39.

Gibbons, J. (2013). The Norm of Belief. Oxford University Press.

Goldberg, S. (1997). Self-ascription, Self-knowledge, and the Memory Argument. Analysis 57 (3):211-219.

Goldberg, S. (1999). The relevance of discriminatory knowledge of content. Pacific Philosophical Quarterly, 80, 136–156.

Goldberg, Sanford C. (2000). Externalism and authoritative knowledge of content: A new incompatibilist strategy. Philosophical Studies 100 (1):51 - 79.

Goldberg, S. (2006). Brown on self-knowledge and discriminability. Pacific Philosophical Quarterly, 87, 301–314.

Goldberg, S. (Ed.). (2015). Externalism, self-knowledge, and skepticism. Cambridge: Cambridge University Press.

Goldman, A. (1976), Discrimination and Perceptual Knowledge, The Journal of Philosophy, 73(20): 771–791.

Goldman, A. (1986). Epistemology and cognition. Cambridge, MA: Harvard University Press.

Goldman, A. (1993). The psychology of folk psychology. Behavioral and Brain Sciences 16: 15–28.

Goldman, A. (1999). Internalism exposed. The Journal of Philosophy, 96(6), 271–293.

Gomes, A. & Parrott, M. (2021). On Being Internally the Same. In Oxford Studies in Philosophy of Mind Volume 1. Oxford: Oxford University Press.

Gopnik, A. (1983). How We Can Know Our Minds: The Illusion of First Person Knowledge of Intentionality. Brain and Behavioral Science, 16: 1–14.

Greco, D. (2014). Could KK Be OK? Journal of Philosophy 111 (4):169-197.

Greco, J. (2005). Justification is not internal. In Steup Matthias & Sosa Ernest (eds.), Contemporary Debates in Epistemology. Blackwell. pp. 257–269.

Hardin, C. L. (1988): ‘Phenomenal colours and sorites’, Nous, 22.

Hawthorne, J.; Rothschild, D. & Spectre, L. (2016). Belief is weak. Philosophical Studies 173 (5):1393-1404.

Heal, J. (1998). Externalism and memory. Proceedings of the Aristotelian Society, 72, 95–110.

Heil, J. (1988). Privileged Access. Mind 386: 238–251.

Helton, G. (2020). If You Can't Change What You Believe, You Don't Believe It. Noûs 54 (3):501-526.

Hinton, J. (1967). Visual experiences. Mind 76 (April):217-227.

Hofmann, F. (2009). Introspective self-knowledge of experience and evidence. Erkenntnis, 71, 19–34.

Horgan, T. & Tienson, J. (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. In David J. Chalmers (ed.), Philosophy of Mind: Classical and Contemporary Readings. OUP USA. pp. 520-533.

Huemer, M. (2006). Phenomenal Conservatism and the Internalist Intuition. American Philosophical
Quarterly 43, 147-158.

Huemer, M. (2007), Compassionate Phenomenal Conservatism, Philosophy and Phenomenological Research, 74(1): 30–55.

Jackson, F. (1994), Armchair Metaphysics, in Meaning in Mind, M. Michael and J. O'Leary-Hawthorne (eds.), Dordrecht: Kluwer Academic Publishers, pp. 23–42.

Jackson, F. (1998), From Metaphysics to Ethics: A Defence of Conceptual Analysis, Oxford: Oxford University Press.

Jackson, F. (2003): Representation and Narrow Belief, Philosophical Issues, 13: 99–112.

Jackson, F. & Pettit, P. (1996). Causation in the Philosophy of Mind. In Andy Clark & Peter Millican (eds.), Philosophy and Phenomenological Research. Clarendon Press. pp. 195-214.

Jaster, R. (2020). Agents' Abilities. Berlin, New York: De Gruyter.

Johnson-Laird, P., & Wason, P. (1970). A theoretical analysis of insight into a reasoning task. Cognitive psychology 1.2: 134-148.

Kahneman, D. (2011), Thinking, Fast and Slow, New York: Farrar, Straus and Giroux.

Kaplan, D. (1989). Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals,' in J. Almog, J. Perry, and H. Wettstein (eds.), Themes from Kaplan, Oxford: Oxford University Press.

Kelly, T. (2008). Evidence: Fundamental concepts and the phenomenal conception. Philosophy Compass 3 (5):933-955.

Kiesewetter, B. (2017). The Normativity of Rationality. Oxford: Oxford University Press.

Kim, J. (1990). Supervenience as a philosophical concept. Metaphilosophy 21 (1-2).

Kim, J. (ed.) (2002). Supervenience. Ashgate.

Kornblith, H. (1988). How Internal Can You Get? Synthese 74(3): 313--327.

Kripke (1980), Naming and Necessity, Cambridge, MA: Harvard University Press.

Lasonen-Aarnio, M. (2010). Unreasonable Knowledge in Philosophical Perspectives, 24(1):1–21.

Lehrer, K., (1990). Theory of Knowledge, first edition, Boulder: Westview Press.

Lepore, E., and Barry L. (1986). Solipsist Semantics, Midwest Studies in Philosophy, 10: 595–614.

Lewis, D. (1966). An Argument for the Identity Theory. Journal of Philosophy 63 (1):17-25.

Lewis, D. (1970), How to Define Theoretical Terms, Journal of Philosophy, 67: 427–46.

Lewis, D. (1972), Psychophysical and Theoretical Identifications, Australasian Journal of Philosophy, 50: 249–258.

Lewis, D. (1975). Adverbs of Quantification. In E. L. Keenan (Ed.) Formal Semantics of Natural Language, (pp. 178–188). Cambridge University Press.

Lewis, D. (1979), Attitudes De Dicto and De Se, Philosophical Review, 88: 513–543.

Lewis, D. (1981), Index, Context, and Content, in Philosophy and Grammar, S. Kanger and S. Ohlman (eds.), Dordrecht: Reidel, pp. 79–100.

Lewis, D. (1996), Elusive Knowledge, Australasian Journal of Philosophy, 74: 549–567.

Littlejohn, C. (2009). The New Evil Demon Problem. Internet Encyclopedia of Philosophy.

Loar, B. (1988). Social Content and Psychological Content, in R. Grimm and D. Merrill (eds.), Contents of Thought, Tucson: University of Arizona Press.

Loar, B. (2003). Phenomenal Intentionality as the Basis of Mental Content, in Martin Hahn and Bjørn Ramberg (eds.), Reflections and Replies: Essays on the Philosophy of Tyler Burge (Cambridge, MA: MIT Press), 229–58.

Loets, A. (2022). Choice Points for a Theory of Normality. Mind 131 (521):159-191.

Lowe, E. J. (2006). Non-cartesian substance dualism and the problem of mental causation. Erkenntnis 65 (1):5-23.

Ludlow, P. (1995). Social externalism, self-knowledge and memory. Analysis, 55(3), 157–159.

Ludlow, P. (1997). On the relevance of slow switching. Analysis 57 (4):285-86.

Lycan, W. (1996). Consciousness and experience. Cambridge, MA: MIT Press.

Madison, B. (2009). On the compatibility of epistemic internalism and content externalism. Acta Analytica, 24, 173–183.

Martin, M.G.F. (1997), The Reality of Appearances, in Thought and Ontology, eds. Sainsbury, Mark. Milan: Franco Angeli: 81-106.

Martin, M.G.F. (2002). The transparency of experience. Mind and Language 17 (4):376-425

Martin, M.G.F.(2004). The Limits of Self-Awareness. Philosophical Studies 120 (1-3): 37-89.

Martin, Michael G. F. (2006). On being alienated. In Tamar S. Gendler & John Hawthorne (eds.), Perceptual Experience. Oxford University Press.

McDowell, J. (1977). On the sense and reference of a proper name. Mind, 86, 159–185.

McDowell, J. ([1983]/(1998)), Criteria, Defeasibility and Knowledge, Proceedings of the British Academy, 68: 455–79, reprinted in McDowell, John (1998). Meaning, Knowledge, and Reality. Harvard University Press: pp. 369-395.

McDowell, J. ([1984]/(1998)). De re senses. Philosophical Quarterly 34 (136):283-294; reprinted in McDowell, John (1998). Meaning, Knowledge, and Reality. Harvard University Press: pp. 214-228.

McDowell, J. ([1986]/(1998)) Singular thought and the extent of inner space. In J. McDowell & P. Pettit (Eds.), Subject, thought, and context (pp. 137–168). Oxford: Oxford University Press.

McDowell, J. (1998). Meaning, Knowledge, and Reality. Harvard University Press.

McHugh, C. (2010). Self-Knowledge and the KK principle. Synthese, 173(3), pp. 231–257.

McCain, K. (2016). Evidentialism and epistemic justification. London: Routledge

McGinn, C. (1984). The Concept of Knowledge. Midwest Studies I nPhilosophy 9. Page references from Knowledge and Reality, C. McGinn, 7–35. Oxford University Press: Oxford.

McKinsey, M. (1991). Anti-individualism and privileged access. Analysis 51 (1):9-16.

McLaughlin, B., & Tye, M. (1998). Is content-externalism compatible with the privileged access? Philosophical Review, 107, 349–380.

Mendola, J. (2008). Anti-Externalism. Oxford University Press, Oxford.

Mendolovici, A. (2018), The Phenomenal Basis of Intentionality, New York: Oxford University Press.

Millar, A. (2019). Knowing by Perceiving. Oxford: Oxford University Press.

Millikan, R.(1993). White Queen Psychology and Other Essays for Alice, Cambridge, MA: MIT Press.

Mitova, V. (2015). Truthy psychologism about evidence. Philosophical Studies 172 (4):1105-1126.

Moran, R. (2001). Authority and Estrangement. Princeton, NJ: Princeton University Press.

Moran, A. (2019). Naïve Realism, Hallucination, and Causation: A New Response to the Screening Off Problem. Australasian Journal of Philosophy 97 (2):368-382.

Morvarid, M. (2015). The epistemological bases of the slow switching argument. European Journal of Philosophy, 23, 17–38.

Morvarid, M. (2019). A new argument for the incompatibility of content externalism with justification internalism. Synthese:1-21.

Moser, P., (1989). Knowledge and Evidence, Cambridge: Cambridge University Press.

Nagel, J. (2013). Knowledge as a Mental State. Oxford Studies in Epistemology 4:275-310.

Nagel, J. (2014). Knowledge: A Very Short Introduction. Oxford University Press.

Neta, R. (2016). Access Internalism and the Guidance Deontological Conception of Justification. American Philosophical Quarterly 53 (2):155-168

Neta, R. & Pritchard, D. (2007). McDowell and the new evil genius. Philosophy and Phenomenological Research 74 (2):381–396.

Nichols, S., and S. Stich. (2003). Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford: Oxford University Press.

Nozick, R., (1981). Philosophical Explanations, Cambridge: Cambridge University Press.

Nuccetelli, S. (Ed.). (2003). New essays on semantic externalism and self-knowledge. Cambridge, MA: MIT Press.

Owens, J. (1989). Contradictory belief and cognitive access. In P. French, T. Uehling, & H. Wettstein (Eds.), Midwest studies (14), contemporary perspectives in the philosophy of language II. Notre Dame: University of Notre Dame Press.

Pautz, A. (2011). Can Disjunctivists Explain Our Access to the SensibleWorld?, Philosophical Issues, 21/1: 384—433.

Peacocke, C., (1983). Sense and Content, Oxford: Oxford University Press.

Perry, J., (1979), The Problem of the Essential Indexical, Noûs, 13: 3–21.

Perry, J. (2001). Knowledge, possibility, and consciousness. Cambridge, MA: MIT Press.

Pettit, P. & Smith, M. (1996). Freedom in belief and desire. Journal of Philosophy 93 (9):429-449.

Pitt, D. (2004). The phenomenology of cognition, or, what is it like to think that P? Philosophy and Phenomenological Research, 69(1): 1–36.

Plantinga, A. (1993), Warrant and Proper Function, Oxford: Oxford University Press.

Pollock, John L. & Joseph Cruz, (1999). Contemporary Theories of Knowledge, 2nd edition, Lanham, MD: Rowman and Littlefield.

Preckel, K., Kanske, P., & Singer, T. (2018). On the interaction of social affect and cognition: empathy, compassion and theory of mind. Current Opinion in Behavioral Sciences, 19, 1-6.

Pritchard, D. (2012). Epistemological Disjunctivism. Oxford University Press.

Pritchard, D., & Kallestrup, J. (2004). An argument for the inconsistency of content externalism and epistemic internalism. Philosophia, 31, 345–354.

Pritchard, D. (2005). Epistemic Luck, Oxford: Clarendon Press.

Pryor, J. (2000) The Skeptic and the Dogmatist, Nous 34: 517-549.

Pryor, J. (2001). Highlights of recent epistemology. British Journal for the Philosophy of Science, 52, 95–124.

Putnam, H. (1967). The Nature of Mental States. In W.H. Capitan & D.D. Merrill (eds.), Art, Mind, and Religion. Pittsburgh University Press. pp. 1-223.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.), Mind, language, and reality: Philosophical papers (Vol. 2, pp. 215–271). Cambridge: Cambridge University Press.

Putnam, H. (1981). Brains in a Vat, in Reason, Truth, and History, Cambridge: Cambridge University Press, Chapter 1: 1–21.

Putnam, H. (1999). The Threefold Cord: Mind, Body and World. New York: Columbia University Press.

Robinson, H. (1985). The General Form of the Argument for Berkeleian Idealism, in Essays on Berkeley: A Tercentennial Celebration, edited by J. Foster and H. Robinson. Oxford: Clarendon Press.

Russell, B.(1905). On Denoting, Mind, 14: 479–493.

Russell, B. (1910). Knowledge by Acquaintance and Knowledge by Description", Proceedings of the Aristotelian Society, 11: 108–28.

Ryle, G. (1949). The Concept of Mind. London: Hutchinson.

Salmon, N. (1989). Tense and Singular Propositions, in J. Almog, J. Perry, and H. Wettstein (eds.), Themes from Kaplan, Oxford: Oxford University Press.

Sainsbury, R.M. (1993): Russell on Names and Communication in Irvine, A. & Wedeking, G. (eds.), Russell and Analytic Philosophy. Toronto: Toronto University Press.

Sainsbury, R.M. (2006). Austerity and Openness. McDowell and His Critics, edited by Cynthia Macdonald and Graham Macdonald, Blackwell Pub.

Sawyer, S. (1999). An Externalist Account of Introspective Knowledge. Pacific Philosophical Quarterly 80 (4):358-378.

Sawyer, S. (2014) Contrastive self-knowledge. Social Epistemology, 28 (2). pp. 139-152.

Sawyer, S. (2015). Contrastive self-knowledge and the McKinsey paradox. In Sanford Goldberg (ed.), Externalism, Self-Knowledge, and Skepticism: New Essays. Cambridge, UK: pp. 75-93.

Schoenfield, M. (2015). Internalism without Luminosity, in Philosophical Issues 25 (1):252-272.

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. Philosophical Review 117 (2): 245-273.

Segal, G. (2000). A Slim Book About Narrow Content. MIT Press.

Shoemaker, S. (1982), Functionalism and Qualia, in Identity, Cause and Mind: Philosophical Essays, Cambridge: Cambridge University Press, 1982.

Shoemaker, S. (1998). Two cheers for representationalism. Philosophy and Phenomenological 58(3), 671-678

Siegel, S. (2004). Indiscriminability and the phenomenal. Philosophical Studies 120 (1-3):91-112.

Siegel, S. (2008) The Epistemic Conception of Hallucination. In Adrian Haddock & Fiona Macpherson (eds.), Disjunctivism: Perception, Action and Knowledge. Oxford University Press. pp. 205–224.

Siewert, Charles P. (1998). The Significance of Consciousness (Princeton: Princeton University Press).

Silins, N. (2005). Deception and evidence. Philosophical Perspectives 19 (1):375–404.

Silins, N. (2020). The Evil Demon Inside. Philosophy and Phenomenological Research 100 (2):325-343.

Smith, M. (2017). The Cost of Treating Knowledge as a Mental State. In A. Carter, E. Gordon & B. Jarvis (eds.), Knowledge First, Approaches to Epistemology and Mind. Oxford University Press. pp. 95-112.

Smithies, D. (2012). Mentalism and epistemic transparency. Australasian Journal of Philosophy, 90(4), 723 742.

Smithies, D. (2019). The Epistemic Role of Consciousness. New York, USA: Oxford University Press.

Smithies, D, and Stoljar, D. (2012). Introspection and Consciousness: An Overview. In Introspection and Consciousness, by Declan Smithies and Daniel Stoljar, Oxford University Press.

Sosa, E. (1999). Scepticism and the internal/external divide. In J. Greco & E. Sosa (Eds.), The Blackwell guide to epistemology (pp. 145–157). Oxford: Blackwell.

Soteriou, M. (2013). The Mind's Construction: The Ontology of Mind and Mental Action. Oxford University Press.

Speaks, J. (2015). Is Phenomenal Character Out There in the World? Philosophy and Phenomenological Research 91 (2):465-482.

Srinivasan, A. (2015). Normativity without Cartesian Privilege, in Philosophical Issues 25 (1):273-299.

Stalnaker, R. (1978), Assertion, Syntax and Semantics, 9: 315–332.

Stalnaker, R.,(1999), Context and Content, Oxford: Oxford University Press.

Stalnaker, R., (2006). On Logics of Knowledge and Belief. Philosophical Studies 128 (1):169-199.

Steup, M. (1999). A Defense of Internalism. In L. Pojman (ed.), The Theory of Knowledge: Classical and Contemporary Readings, 2nd edition. Wadsworth Publishing.

Strawson, G. (1994). Mental Reality. MIT Press.

Sturgeon, S. (2008a). Reason and the grain of belief. Noûs 42 (1):139–165.

Sturgeon, S. (2008b) Disjunctivism About Visual Experience, in Disjunctivism: Perception, Action, Knowledge, eds. MacPherson, Fiona and Haddock, Adrian. Oxford: Oxford University Press: 113—43.

Sutton, J. (2005). Stick To What You Know. Nous 39: 359-96.

Sutton, J. (2007). Without Justification. (Cambridge, MA: MIT University Press).

Tillman, C. (2012). Reconciling justificatory internalism and content externalism. Synthese, 187, 419–440.

Tye, M. (1995a). Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. MIT Press.

Tye, M.(1995b). What what its like is really like. Analysis 55 (2):125-126.

Tye, M. (1998). Externalism and memory. Proceedings of the Aristotelian Society, 72, 77–94.

Tye, M. (2009). Consciousness Revisited: Materialism Without Phenomenal Concepts. MIT Press

Vahid, H. (2003a). Content externalism and the internalism/externalism debate in justification theory. European Journal of Philosophy, 11(1), 89–107.

Vahid, H. (2003b). Externalism, slow-switching, and privileged self-knowledge. Philosophy and Phenomenological Research, 66(2), 370–388.

Warfield, T. A. (1997). Externalism, privileged self-knowledge, and the irrelevance of slow switching. Analysis 57 (4):282-284.

Wedgwood, R. (2002). Internalism explained. Philosophy and Phenomenological Research, 66, 349–369.

Wedgwood, Ralph (2014). Rationality as a Virtue. Analytic Philosophy 55 (4):319-338.

Wedgwood, R. (2017). The Value of Rationality. Oxford University Press.

Williamson, T. (1990). Identity and Discrimination. Wiley-Blackwell.

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

Williamson, T. (2007). On being justified in one's head. In M. Timmons, J. Greco, & A. Mele (Eds.), Rationality and the Good (pp. 106–122). Oxford: Oxford University Press.

Williamson, T. (2009). Probability and Danger. Amherst Lecture in Philosophy.

Williamson, T. (2020). Justifications, Excuses, and Sceptical Scenarios. In J. Dutant and F. Dorsch, (eds.), The New Evil Demon. Oxford: Oxford University Press. Archived in Phil

Whittle, A., 2010, Dispositional Abilities, Philosophers' Imprint, 10(12): 1–23.

Yablo, Stephen (1992). Mental causation. Philosophical Review 101 (2):245-280.

Yli-Vakkuri, J. & Hawthorne, J. (2018): Narrow Content. Oxford University Press.